Faculty of Psychology and Educational Sciences
Subfaculty of Psychology and Educational Sciences – KULAK
Interactive Technologies (ITEC- iMinds)
Methodology of Educational Sciences Group

Three-level synthesis of single-subject experimental data

Further developments, empirical validation and applications

Mariola Moeyaert

Supervisor: Prof. dr. Wim Van den Noortgate

Co-supervisor: Prof. dr. John Ferron

Co-supervisor: Prof. dr. Patrick Onghena

Dissertation offered to obtain the degree of

Doctor of Educational Sciences (PhD)

2014

Guidance Committee
Prof. dr. Wim Van den Noortgate (supervisor)
KU Leuven, Faculty of Psychology and Educational Sciences
Prof. dr. John Ferron (co-supervisor)
University of South Florida, Department of Educational Measurement & Research
Prof. dr. Geert Molenberghs
KU Leuven, Department of Public Health and Primary Care
Prof. dr. Betsy McCoach
University of Connecticut, Department of Educational Psychology
Prof. dr. Ralph Schlosser
Northeastern University, Department of Speech Language Pathology and Audiology
Dr. Mieke Heyvaert
KU Leuven, Faculty of Psychology and Educational Sciences

Examination Committee
Prof. dr. Marc Depaepe (chair)
KU Leuven, Faculty of Psychology and Educational Sciences
Prof. dr. Wim Van den Noortgate (supervisor)
KU Leuven, Faculty of Psychology and Educational Sciences
Prof. dr. John Ferron (co-supervisor)
University of South Florida, Department of Educational Measurement & Research
Prof. dr. Patrick Onghena (co-supervisor)
KU Leuven, Faculty of Psychology and Educational Sciences
Prof. dr. Geert Molenberghs
KU Leuven, Department of Public Health and Primary Care
Prof. dr. David Rindskopf
City University of New York, Department of Educational Psychology
Prof. dr. Eva Ceulemans
KU Leuven, Faculty of Psychology and Educational Sciences

First edition, first print March, 2014

Copyright © 2014 by Mariola Moeyaert
ISBN:
Printed by University Press, Zelzate
Cover designed by Gerdy Vandermeersch

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without written permission of the author.

Three-level synthesis of single-subject experimental data: Further extensions, empirical validation and applications.

Dra. Mariola Moeyaert

Supervisor: Prof. dr. Wim Van den Noortgate,

Co-supervisor: Prof. dr. John. M. Ferron, and Prof. dr. Patrick Onghena.

During the last decade, there is a growing interest in using single-subject experimental designs (SSED) in a variety of different research fields in education as a means to investigate the effectiveness of one or multiple treatments (Barlow, Nock, & Hersen, 2009; Morgan & Morgan, 2001). In an SSED study, one or a few subject(s) (or another entity) is the focus of interest and is measured repeatedly during successive conditions, usually a baseline condition (in which no treatment is present) and a treatment condition (Barlow et al., 2009; Kazdin, 2011; Onghena, 2005). By comparing scores from both kinds of conditions, a single-case researcher can assess the functional relationship between the condition and the outcome scores on the dependent variable (e.g., the score on a statistical test). Although SSEDs are growing in popularity and are valued, the external validity is often questioned because of the small number of subjects under investigation in one SSED study. In order to establish an evidence base for treatment effects, several SSED studies can be combined and a three-level data structure becomes visible, namely measurement occasions are nested within subjects and subjects in turn are nested within studies. The synthesis of the studies can inform research, practice and policy and important decisions can be made based on the synthesis results. In this dissertation, we focus on one specific flexible methodological framework that takes this hierarchical data structure into account and that can be used to summarize SSED data across subjects and across studies, namely three-level modeling. This multilevel approach is promising and enables estimating treatment effects across cases and across studies in addition to study-specific and subject-specific treatment effects. Furthermore, variation in these treatment effects between studies and between subjects can be estimated, multiple predictors can be added, autocorrelation and heterogeneous variance can be modeled, etc. This dissertation is comprised of two large parts; a methodological part which is the product of four methodological papers and an applied part in which three applied papers are presented. As a consequence, the purpose of this dissertation is twofold; (1) empirically validate the methodology of multilevel modeling, and (2) enhancing the understanding of this flexible way of synthesizing SSED data and promoting the use of multilevel models by giving practical illustrations**.** In this way, the dissertation is of interest to the methodologist, the single-subject meta-analyst, and the applied single-subject researcher. The methodologist will be challenged to examine suggestions for further research, the meta-analyst will be encouraged to use the multilevel model as it provides a flexible way to model a variety of different SSEDs, and the practical implications will guide applied single-subject researchers in setting up SSED studies, doing the analysis and interpreting and reporting their results. After a general introduction (*Chapter 1*), we focus in *Part 1* on intensive Monte Carlo simulation methods to validate the basic three-level model and some extensions. We start with the empirical validation of the basic multilevel model (*Chapter 2*). A commonly encountered issue when synthesizing SSED studies is standardization which will be the focus of interest in *Chapter 3*. SSEDs are vulnerable to several threats to internal validity. We suggest one way to take external event effects into account (*Chapter 4*). In the last chapter of the first part (*Chapter 5*), we evaluate the consequences of misspecifying the covariance matrix at the second and third level of the multilevel model. This allows examination of the robustness of the three-level model. In the second part (*Part 2*) of this dissertation, we aim to provide a broad understanding of the options, the flexibility and the use of the multilevel modeling framework by giving empirical illustrations using different empirical datasets. In *Chapter 6*, the design matrix specification is elaborated and illustrated using graphical presentations and real datasets. We explain in detail the process from single-level analysis to multilevel analysis of SSED data in *Chapter 7*. In the last chapter (*Chapter 8*), we illustrate how to combine several types of SSEDs such as simple AB designs, multiple-baseline designs, ABAB reversal designs and alternating treatment designs using one multilevel modeling framework on a real dataset. A third part (*Part 3*) of this dissertation is comprised of two chapters. *In Chapter 9*, a summary of the main findings is given, and methodological issues and implications for further research are discussed. We end this dissertation by giving suggestions for further research (*Chapter10*).

Drie-niveau synthese van single-subject experimentele data: Verdere uitbreidingen, empirische validatie en toepassingen.

Dra. Mariola Moeyaert
Promotor: Prof. dr. Wim Van den Noortgate,
Copromotor: Prof. dr. John. M. Ferron, and Prof. dr. Patrick Onghena

Gedurende het laatste decennium is er een groeiende interesse om single-subject experimentele designs (SSED) in verschillende onderzoeksdomeinen in educatie toe te passen om de effectiviteit van één of meerdere behandelingen te onderzoeken (Barlow et al., 2009; Morgan & Morgan, 2001). De interesse van de SSED onderzoeker gaat uit naar één of meerdere individuen (of een entiteit zoals een school) dat geobserveerd en gekwantificeerd wordt gedurende opeenvolgende meetmomenten. Een SSED wordt gekenmerkt door een baseline conditie (waarin men geen behandeling toedient), gevolgd door een behandelingsconditie (Barlow et al., 2009; Kazdin, 2011; Onghena, 2005). Door het vergelijken van baseline- en behandelingsobservaties, kunnen SSED onderzoekers nagaan of er al dan niet een functioneel verband bestaat tussen de conditie en de geobserveerde score (bijvoorbeeld de behaalde score op een statistische test). Niettegenstaande de groeiende populariteit en waardering voor SSEDs, rijzen er vragen betreffende de externe validiteit van de onderzoeksresultaten aangezien slechts een beperkt aantal personen deel uitmaken van de SSED studie. Het drieniveau model kan gebruikt worden om SSED studies samen te vatten wat resulteert in meer extern valide uitspraken met betrekking tot het effect van een behandeling. De synthese van SSED data over subjecten en over studies heen kan onderzoek, praktijk en beleid inspireren en informeren en belangrijke beslissingen kunnen genomen worden op basis van deze resultaten. De drieniveau benadering is veelbelovend en maakt het mogelijk om behandelingseffecten over subjecten en over studies te schatten bovenop studie-specifieke en subject-specifieke behandelingseffecten. Bovendien kan variantie in behandelingseffecten tussen subjecten en tussen studies geschat worden, predictoren kunnen toegevoegd worden, autocorrelatie en heterogene variantie kunnen gemodelleerd worden, enz. Dit proefschrift bestaat uit twee grote delen: een methodologisch en een toegepast gedeelte. In het methodologisch deel wordt het basis drieniveau model en verschillende uitbreidingen gevalideerd resulterende in vier methodologische manuscripten. Het toegepaste gedeelte bestaat uit drie manuscripten waarin toepassingen van het drie-niveau model verhelderd en geïllustreerd worden met behulp van empirische illustraties. Het doel van dit proefschrift is dan ook tweevoudig, enerzijds het empirisch valideren van de multiniveau methodologie, en anderzijds het geven van praktische toepassingen. Op deze manier is dit proefschrift informatief voor zowel de methodoloog, statisticus, meta-analist als de toegepaste onderzoeker. De methodoloog zal uitgedaagd worden om verder op zoek te gaan naar antwoorden op onopgeloste vragen en kan de suggesties voor verder onderzoek onder de loep nemen. De statisticus en meta-analist worden verder op weg geholpen om inferenties te maken betreffende behandelingseffecten. De praktische toepassingen van het tweede gedeelte zetten toegepaste SSED onderzoekers op weg bij het opzetten van een studie, het uitvoeren van de analyse en het interpreteren en rapporteren van de resultaten. Na een algemene inleiding (*Hoofdstuk 1*), focussen we in het eerste gedeelte op computer-intensieve simulatiestudies om het basis drieniveau model en verschillende uitbreidingen van dit model te onderzoeken. We starten met de empirische validatie van het basis drieniveau model (*Hoofdstuk 2)*. Een vaak voorkomend probleem bij het combineren van SSED studies is standaardisatie wat de onderzoek focus is in *Hoofdstuk 3***.** SSEDs zijn gevoelig voor verschillende bedreigingen aan interne validiteit en daarom stellen we een mogelijkheid voor om externe factoren in rekening te houden (*Hoofdstuk 4*). In *Hoofdstuk 5* evalueren we de robuustheid van het drieniveau model. Het doel van het tweede gedeelte van dit proefschrift is het verstrekken van de nodige informatie en richtlijnen om SSED data samen te vatten gebruik makende van het drieniveau model. In *Hoofdstuk 6* leggen we de nadruk op de specificatie van de SSED matrix. In een volgend hoofdstuk**,** *Hoofdstuk 7*, leggen we stap per stap het proces uit om van een single-niveau analyse over te gaan naar een multiniveau analyse. In het laatste hoofdstuk van het tweede deel, *Hoofdstuk 8*, illustreren we hoe verschillende SSED types (AB-fase designs, multiple-baseline designs, reversal designs, and alternating treatment designs) gecombineerd kunnen worden gebruik makende van één multiniveau analyse. Tot slot geven we in een derde gedeelte een samenvatting van de belangrijkste onderzoeksresultaten, presenteren we beperkingen en geven we implicaties (*Hoofdstuk 9*). Verschillende suggesties voor verder onderzoek worden gepresenteerd in *Hoofdstuk 10*.

Acknowledgements

Is my doctoral research project really coming to an end after two and a half years of conducting research, and are these my final words of thanks? No it is not, my research career has only just begun. Therefore, I hope I can keep on counting on everyone who has supported me during the last couple of years.

My friends, family and boyfriend have the impression that I might be some sort of lonely researcher, analyzing simulated data all day and night, and that I live on an isolated planet. To be honest, they might be right. Sometimes they succeed in putting both of my feet back on the ground, but from time to time I prefer staying on my own planet, with my own thoughts and models. For me, it was sometimes hard to find a good balance between research and relaxation. In the long run research won, which made time with family, friends and especially my boyfriend very precious. I sincerely apologize to everyone for whenever I did not have time for them.

Both my entourage on the one hand, and the research center and especially my main supervisor Wim Van den Noortgate on the other hand, were needed to be able to succeed and end here, as a Doctor. To express it in multilevel modeling terms: this doctoral dissertation, at the highest level, is supported by two bigger levels: the level of the research team and on the base a bigger level: my family, friends and especially my boyfriend, Tom Verhack. I am proud to present here the product of years of conducting research, so to all of you who contributed to this product: many congratulations and many thanks. I hope you enjoy the reading of a work that has been inspired by you.

First of all, I want to thank Prof. dr. Wim Van den Noortgate for giving me the great opportunity to start this *PhD*. I remember our first conversation about multilevel modeling of single-cases as if it was yesterday. You explained me the multilevel model, single-cases and simulation studies. I did not have any background in these research fields, but you believed in my capacities, my ambitions and my motivations to successfully start this *PhD*. Wim, you are the pioneer in the field of multilevel modeling and thanks to you I could further study this promising method and further optimize the models. I have learned a lot from you and you have pushed me to the edge of my own capacities and challenged me to go beyond the scope of my own knowledge. In the beginning I did not realize what great professor, mentor, and person you were. You let me choose my own research path independently and this allowed me to tackle my own issues and difficulties. Although this sometimes resulted in a waste of

time and frustrations, I realize that this method was the best learning school. And if I was completely lost, you were there to put me back on track. During the last years, we wrote some grant applications (i.e., junior mobility program, Flemish foundation, and conference traveling grants) and a number of articles. Sometimes I questioned their success rate but you always believed in my work and put high hopes in it and you were always right. Sometimes I had hard times to deal with negative feedback and revisions and I was dispirited. You taught me to transform negative feedback into learning opportunities and in chances to enhance my research quality. In addition to being a great professor, you also have a great personality, which is confirmed by the students. Although we worked a lot during the last years, we also had some time for pleasure: the weekend in the Ardennes, the Ypres Tours, free time at conferences, etc. Wim, it was a pleasure collaborating with you and I hope to continue working with you in future!

Another great person I want to thank is my co-supervisor Prof. dr. John Ferron from the University of South Florida (*USF*). John, you are the person I am looking up to. First of all, I am impressed by your numerous publications with experts in the field of single-subject research and your in-depth knowledge about these type of designs. This was extremely helpful for my dissertation. I also appreciate the quick feedback you provided to my work, which proves that distance is really relative. To my knowledge I am not aware of any professor who makes that much time for his or her students and gives such quick and detailed feedback. To be honest, my English writing skills were poor, but they improved a lot thanks to you and this in turn increased the quality of my research articles. I am also very thankful for having been invited to work at *USF* for seven months. This enhanced my research skills and I could learn a lot from you and your *PhD* students. It was also a great opportunity to develop my research network. By following the course you gave about single-subjects at *USF*, I realized that all students were big fans of you, and that you are a very motivating and inspiring person, which I completely confirm. John, thank you for believing in me, for giving me first authorship for the special issue of the Journal of School Psychology, for working together on the between-within paper, and for giving co-authorship of a special issue for the journal of Neuropsychological Rehabilitation. I have to thank you in so many ways. I sincerely hope that we can continue to collaborate in the future.

Another large contributor to this doctoral dissertation is Prof. dr. Tasha Beretvas from the University of Texas (*UT*). Although you are not one of the co-supervisors of this dissertation, I consider you as one. Your in-depth comments on my research papers forced me to go beyond the scope of commonly accepted issues and optimized the research content of my papers. I am also grateful for the grammar and spelling edits to my research papers. In addition to this, you are a very motivating person, and always in a good mood. By working closely together at *UT*, I realized that work and pleasure can be melted into one. Thank you so much for offering funding for my research stay at *UT,* which was unfortunately too short. This enabled me to attend the Texas Universities' Educational Statistics and Psychometrics Meeting you organized, and to work closely together with you and your *PhD* students, which were great learning opportunities. Thank you for being such a good mother to me while being in Austin, for your encouraging feedback, for your kind words and detailed comments to my research papers. You are one of the most enthusiastic and open researchers I know.

Prof. dr. Patrick Onghena, as founding father and pioneer of single-case designs, from the beginning onwards you were interested in this doctoral project, which meant a lot to me. Although you were not a co-supervisor at the beginning, I was grateful to add you as one thanks to the *FWO* (Flemish Foundation) funding. Your comments during seminars and encouraging emails always motivated me to continue discovering more about single-subject designs. For instance, one comment you gave about the interdependence between subjects in a multiple-baseline design resulted in a research manuscript about external event effects. Also, thank you for sending important published articles about the analysis and meta-analysis of single-cases. I also want to thank you for the recommendation letters you wrote for the *FWO*, the Belgian American Educational Foundation (*BAEF),* and the Junior Mobility Program (*JUMO)*. Also, thanks to you, I got the opportunity to give a seminar at the University of Delaware which resulted in a collaboration with a very interesting research group. Thank you so much, Patrick, it has been a pleasure working with you and I hope I can still count on you in the future.

Other two important contributors to this dissertation are Maaike Ugille and Mieke Heyvaert from the Methodology of Educational Science Research Center. Maaike, we started a *PhD* around the same date and we could tackle a lot of challenges together and motivate each other to continue doing so. Mieke, your knowledge about applied single-case research was extremely helpful and I am so thankful for the large datasets you provided, which we could use to illustrate our multilevel modeling methods. Maaike and Mieke, I appreciate both of you as researchers and as people, and I learned a lot from our collaboration, thank you!

I would also like to thank the other colleagues and ex-colleagues from the KU Leuven-Kulak and the Methodology of Educational Science research center for the relaxing conversations during breaks and lunches. In particular, I would like to thank Chloé Meredith and Giovanni Delaere, two ex-class mates, very close friends and meanwhile colleagues for their support and sincere interest in my research projects from the beginning onwards. I had so much fun and distractions with you, but you also asked a lot of questions about my research and how it was going. We went through milestones in life together and I cannot reword what that means, but it resulted in unconditional friendship and support. Giovanni, I am so grateful that you, as outsider and expert in the English language volunteered to read this dissertation in detail and to correct for spelling and grammar. Another great colleague-friend is Niki Mistiaen (research assistant at the Kulak). Thank you, Niki, for the relaxing swimming moments and the confidential talks. We are a good match!

I want to thank the members of my Guidance Committee: Mieke Heyvaert, Ralph Schlosser, Betsy McCoach and Geert Molenberghs. You were the perfect mix of applied single-case researchers and methodologists, which was needed because my doctoral dissertation is purposed for both type of research groups. Without your valuable comments and input to my preliminary doctoral text, this dissertation would not have been optimized.

I would also like to thank Prof. dr. David Rindskopf (City University of New York), Prof. dr. Eva Ceulemans (KU Leuven), Prof. dr. Geert Molenberghs (KU Leuven) and Prof dr. Marc Depaere (Vice rector Kuleuven @ Kulak) for being a member of the Examination Committee. David, I am impressed by your expertise in the context of SSEDs, and I am convinced that you will ask me challenging questions during my defense, but these will all be great learning opportunities. In the future, I hope to work more closely with you, because I am interested in your advanced Bayesian estimation methods, which seems to be promising in contexts of analyzing single-subject data. Eva, you are probably less familiar with multilevel modeling and single-cases, but this makes you a very valuable member who can criticize the multilevel modeling method to synthesize single-subject data as an outsider. Geert, it is a pleasure to have you as member of my Examination Committee because you are familiar with both multilevel modeling and repeated measures, but in another research domain, namely Biomedics. You were also member of my Guidance Committee and you gave me very interesting and encouraging comments which motivated me to broader my research view and think more interdisciplinary . Thank you for that. Marc, it is an honor to have you as chair of my Examination Committee. When I started my studies five years ago at the Kulak, I was impressed by the course ‘History of Education’, and three years ago I could join you as

colleague at the subfaculty of Psychology and Educational Sciences. Even when you became vice-rector of the KU Leuven @ Kulak, you stayed involved with the subfaculty and interested in my research work. You also showed a lot of interest in my political career (as councillor). Thank you for your time and interest.

Next, I want to thank the funding organizations, because without their financial support, this dissertation would not have been possible. From August 2011 to September 2013, I was funded by the Institute of Educational Sciences (*IES*, Grant number R305D110024), after which I was funded by the *FWO*. I am also grateful for the support of the *JUMO* which made my international stay at *USF* and *UT* possible.

Of course I want to thank my parents for the financial support for completing my Bachelor and Master's program. Thank you parents and parents-in-law for helping me with my housekeeping, asking how my research is going, being proud of me, taking me out and supporting me with my research (without really knowing what I am doing). This meant a lot to me. I would also like to mention my grandparents, who never knew what my work was really about and still think that I am becoming a doctor to heal people. Although you did not say it with that many words, I am sure that you are proud of me. I am also thankful for my brothers, sister, brothers-in law and sisters-in-law, first for being interested (or not) in what I am doing, and second for just being there in case I am in trouble. I would like to thank Ben Verhack in particular for his motivating words, helping hands, and interest. During my international stay abroad, I also created a new family. Patricia, Ismael, Ismael and Patti, thank you so much for the seven great months in Tampa, for driving me to the university, for the interesting talks about doing research in the States and for being so interested in my research. You also showed me all the great places in Florida and I am very thankful for that.

My friends mean a lot to me and are like my real family. I would like to thank some of them in particular. Through all stages in life, I met some people, lost some people, but the real friends stayed. First of all, Stefanie Kestelyn, I have known you since we were little kids. I was your maid of honor and I will be the godmother of your first child. We have shared so many great things together, but we have also supported each other during hard times. Also Charlotte Candry, you were there through all the stages in my life, like meeting our first boyfriend together. And thank you for all the conversations and crazy distractions. I would also like to thank Lien Vandelanotte, we met in secondary school and ever since then, we were inseparable. Thank you for all the great conversations, crazy parties and times that I could stay at your place to catch trains to go to the University of Leuven. When I started my Bachelor, I shared an apartment with Nele Demuynck, Sofie Delmulle, and Petra

Deleersnyder. We are study soulmates, know a lot of each other and I enjoy all our reunions. I am also grateful for meeting other people who were interested in studying mathematics during my Bachelors program and especially Joni Deman. Thanks to all my study friends for always being that interested in my research work. During the last stage of my life, I met Stéphanie Vermeersch, Ine Vandenberghe, and Evelien Covemaeker, who are my travel buddies, friends with who I can relax, make fun and have great conversations.

October 14th, 2012 was an important date as I was elected as councillor of the city of Ieper. I would like to thank the citizens if Ieper and my colleague-politicians for the learning opportunities and the understanding when I missed meetings because of my *PhD*. I would like to thank Jan Durnez and Yves Leterme in particular because they are very motivating and inspiring persons for me. Both are involved in educational policy and show interest in my research topic. Thank you, Jan and Yves, for believing in my capacities, for the support and the great learning opportunities. I hope to count on you in future.

The last person I want to thank is the most important person in my life, and that is my boyfriend Tom Verhack. Tom I don't know where to start, but all I am is because of you. I know words cannot cover what you mean to me and how important your support was during the whole *PhD* process. You were there from the beginning onwards to encourage me, you believe in me and in my capacities, more than I believe in myself. You were the one who convinced me to start this *PhD* and you are the big power behind this dissertation. This *PhD* has been a proof of our unconditional love, as we were separated for seven months and came stronger out of this. You were always there to cheer me up when I had a frustrating or unproductive day, to distract me from work and to take me out to destress. You are the only one who can reveal all my levels. I can't thank you enough for this.

Of course I forgot a lot of people and I sincerely apologize for this. Thanks to everyone for acting like you were interested or for really being interested in my *PhD*.

Mariola Moeyaert

March 2014

Table of contents

Chapter 1|
General Introduction

Background: Funding, Additional Publications and International Collaborations

This doctoral dissertation is funded by the Flemish Foundation (*FWO*, Grant number ZKC6624*)*. and is part of a larger research project funded by the Institute of Educational Sciences (*IES*, Grant number R305D110024). The *IES* project is embedded in an international context and is the result of a strong collaboration between the KU Leuven, the University of South Florida (*USF*) and the University of Texas (*UT*). The topic of the *IES* grant is similar to the topic of this dissertation and investigates extensions to the multilevel modeling of single-subjects. I worked closely together with Maaike Ugille (KU Leuven), who focused in her study on combining effect sizes instead of raw data using the multilevel modeling framework (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2013, Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2014). The second year of my *PhD*, I spent six months at *USF*, to work more closely together with John Ferron, one of my co-supervisors. This international stay resulted in two publications as second author, which are not included in this doctoral dissertation. The first paper deals with explaining the basics of multilevel modeling to an applied audience (Baek, Moeyaert, Petit-Bois, Beretvas, Van den Noortgate, & Ferron, 2013). The second paper is more challenging and presents and validates a between-subject estimator in context of multiple-baseline designs which is resistant against threats to internal validity (e.g., external event effect, Ferron, Moeyaert, Beretvas, & Van den Noortgate, 2014). This latter study is closely related to the study reported in *Chapter 4* entitled 'Modeling external event effects in the three-level analysis of multiple-baseline across participants design'. After staying six months at *USF*, I spent one month at *UT* to collaborate with the other research team involved in the *IES* research grant. Together with Tasha Beretvas and one of her *PhD* students, Rommel Bunuan, I worked on a paper concerning dependent effect sizes in contexts of alternating treatment designs (Moeyaert, Bunuan, & Beretvas, 2014), which we will present at the annual meeting of the American Educational Research Association (2014).

General Introduction: Single-Subject Experimental Designs and Multilevel Modeling

Over the past decade, evidence-based practices and policy explicitly rely on scientific research (National Research Council, 2002, Shadish & Rindskopf, 2007). This resulted in developments such as the What Works Clearinghouse (*WWC*), which main focus is on evaluating the quality of published research and determining the effectiveness of specific practices in educational contexts. Single-subject experimental design (SSED) studies have provided scientifically sound evaluations of treatment effects in a variety of different research fields such as in biomedical research, school effectiveness, behavior modification, school psychology, and special education for more than 50 years (Gast, 2010; Kennedy, 2005, Kratochwill, 1978; Tawney & Gast, 1984; Busse, Kratochwill, & Elliott, 1995; Chorpita, Albano, Heimberg, & Barlow, 1996; Barlow & Hersen, 1984; Kratochwill & Levin, 1992), and are included in the *WWC* single-case design technical documentation guidelines (Kratochwill et al., 2010). In the writing of the *WWC* documentation, a panel of experts in SSED and analysis of SSEDs were gathered to describe SSED studies as scientific evidence available for quantitative synthesis. In addition, increased attention has been placed on SSEDs as the *IES* has included this type of design as a rigorous research design within its research grant framework. Many researchers recognize the valuable contributions SSED methods have made to educational research (e.g., National Research Council, 2002; Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005). A search of the Social Science Citation Index (SSCI) within the Web of Science using the key terms “single-case” or “single-subject” shows an increase in the number of published items over the last decades (see Figure 1.1).

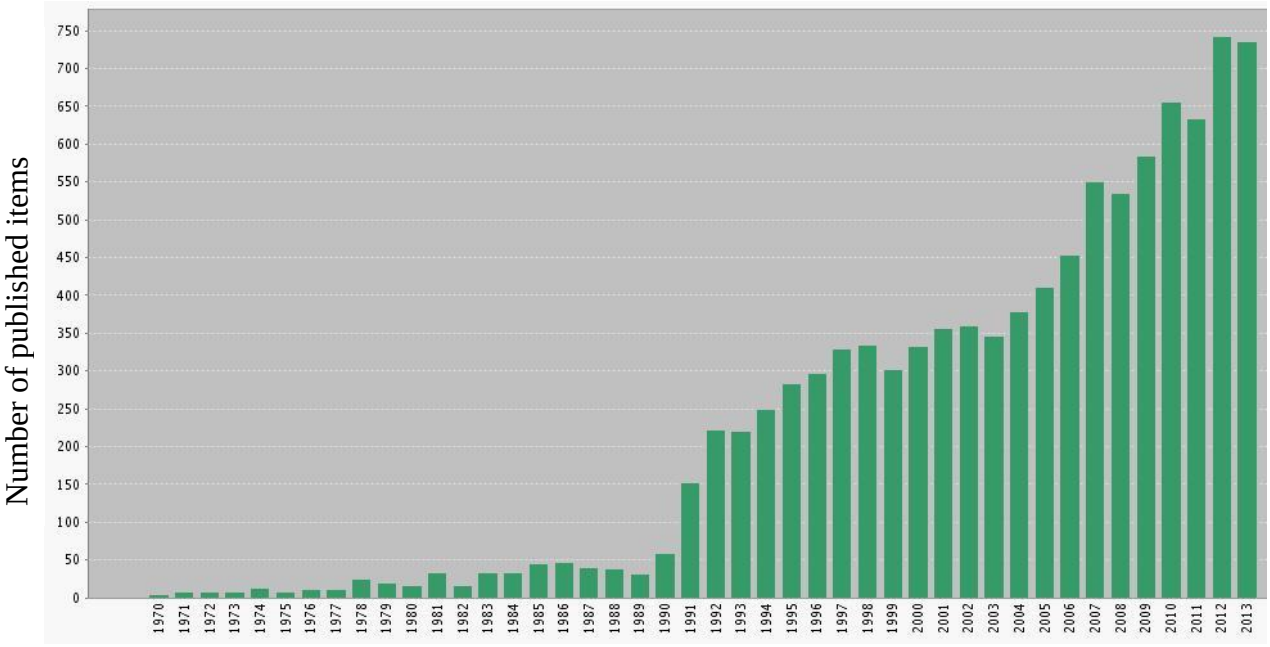

Figure 1.1. Graphical display showing the increase in the number of published items for the keywords "single-case" or "single-subject" between 1970 and 2013 using the Social Science Citation Index within the Web of Sciences.

Because of the popularity of SSEDs within and across a variety of different research fields, a large number of SSEDs is available for quantitative synthesis (Shadish & Rindskopf, 2007). In the remainder of this chapter, we give a brief introduction into SSEDs (i.e., definition, characteristics, and types), describe how an SSED can be analyzed and how the SSED study research findings can be synthesized across SSED studies using the multilevel modeling framework in order to contribute to evidence based research, to inform policy and research, and to improve practice (Shadish & Rindskopf, 2007).

1.1 Single-Subject Experimental Design

1.1.1 Definition

According to the *WWC* standards, an SSED study is identified by three important characteristics: (1) data are gathered, analyzed and interpreted for one entity (this entity can be one participant or a group of participants e.g., a classroom, a school or an organisation), (2) the participant(s) is (are) observed repeatedly during baseline(s) and treatment(s) phase(s), and (3) outcomes during and after the treatment are compared with outcomes prior to treatment (Barlow & Hersen, 1984; Kazdin, 2011 ; Kratochwill et al., 2010; Onghena, 2005). Despite what might be expected from the name "single-subject experimental design study", usually more than one participant (i.e., subject or case) is included in the single-subject

experimental study. The main focus of this design lies in assessing whether there is a causal relation between the introduction of a treatment and the change in a dependent variable (Levin, O'Donnell, & Kratochwill, 2003; Onghena, 2005). This implies that in an SSED study, a case is observed longitudinally under several experimental conditions or phases (at least one baseline condition in which no treatment is given and one treatment or intervention condition). SSED studies provide detailed information about variations in the treatment effect related to specific subjects under investigation. This tends to be lost in group-comparison designs because they only provide averages and effect sizes for the entire group (Barlow & Hersen, 1984). In addition to individual variation, this type of design also allows the individual to be measured at various points in time, thereby allowing the treatment effect to be evaluated with more than a single observation, which allows researchers to see how the treatment effect changes over time (i.e., identifying trends). Due to the fact that in one SSED study only a small number of individuals is needed, researchers are able to study populations of people that have a low prevalence rate (e.g., children with special needs). Another advantage is that these designs reduce the gap between research and practice by allowing practitioners to implement SSEDs in their natural settings (Morgan & Morgan, 2001). Within the SSED, the subject provides its own control for purposes of comparison (Kratochwill et al., 2010; Perone, 1999). For example, the subject's series of outcome variable values prior to the intervention is compared with the series of outcome variable values during (and after) the intervention. In literature, single-subject experimental designs (Guralnick, 1978; McReynolds, & Thompson, 1986) have taken on a variety of different names, such as single-case design (Gingerich, 1984), intrasubject replication design (Gentile, Roden, & Klein, 1972), reversal design (Gentile et al., 1972), individual organism research (Michael, 1974), intrasubject design (Center, Skiba, & Casey, 1985-1986), intrasubject experimental design (White, Rusch, Kazdin, & Hartmann, 1989), N = 1 design (Strube, Gardner, & Hartmann, 1985), N of 1 design (Gorsuch, 1983), one-subject experiment (Edgington, 1980), interrupted time series (Michielutte, Shelton, Paskett, Tatum, & Velez, 2000), and small-n design because some single-case studies investigate more than one subject (Kratochwill & Levin, 1992). In this doctoral dissertation we will use the terms single-case experimental design and single-subject experimental design interchangeably.

The three main characteristics of an SSED study (focus on one entity, repeated measures across time, and experimental control) can be used to situate single-subject research designs in a broader research context, which might help understanding how these designs differ from closely related designs. Similar to group-comparison designs, also in the area of

SSEDs a distinction can be made between experimental and quasi-experimental SSED studies. In experimental studies random assignment of measurement occasions to treatments is feasible (e.g., Bulté & Onghena, 2009; Edgington & Onghena, 2007; Koehler & Levin, 2000; Manolov & Solanas, 2009), whereas this is not the case in quasi-experimental studies. The major difference between a typical SSED study and group-comparison design study is that these latter type of designs focus on average treatment effect estimates, whereas SSED studies focus on a limited number of preselected individuals and subject-specific treatment effect estimates are obtained. In SSEDs, an entity is measured repeatedly across time, whereas group-comparison designs usually incorporate one measurement per subject. This implies that in group-comparison designs, often no within-subject trends can be identified. An SSED should not be confounded with a (qualitative) case study or observational case study research. In a typical case study, a single entity is involved but there is not a purposeful manipulation of an independent variable nor are there necessarily repeated measures. Most case studies are reported in a narrative way while results of SSEDs are presented numerically or graphically. In observational time series research there are also repeated measures but there is an absence of a designed treatment. SSEDs are experimental or quasi-experimental designs (in case there is no random assignment), because they are characterized by the active manipulation of the independent variable by the researcher. SSEDs and longitudinal designs have in common that subjects are measured repeatedly across time allowing identifying trends. In longitudinal designs, subjects are measured across a long period (which can be years or decades), whereas in SSEDs, measurement occasions tend to be closer together in time so that the SSED can be completed in weeks or months. In an SSED, one or multiple subject(s) can be involved, whereas in longitudinal designs, multiple subjects are observed simultaneously across time (Verbeke & Molenberghs, 2009). In this doctoral dissertation, we focus on SSEDs as a means to build further on an evidence base for intervention effects.

1.1.2 Types of single-subject experimental designs

1.1.2.1 Basic single-subject experimental designs

There are several types of SSEDs. The most basic type is an AB, or interrupted time series design (i.e., data are collected repeatedly over time, but the baseline condition is interrupted by a treatment, see Figure 1.2). A basic SSED is characterized by an A-phase (i.e., baseline condition) followed by a B-phase (i.e., treatment or intervention condition). In SSED research there has been a tradition to graphically display the data, such as in Figure 1.2 (Barlow & Hersen, 1984, Kartochwill et al., 2010; Kazdin, 2011). In Figure 1.2, relatively stable outcome scores during a baseline phase are obtained, an increase in outcome scores due to the treatment is observed, and during the treatment the outcome scores gradually decrease across time.

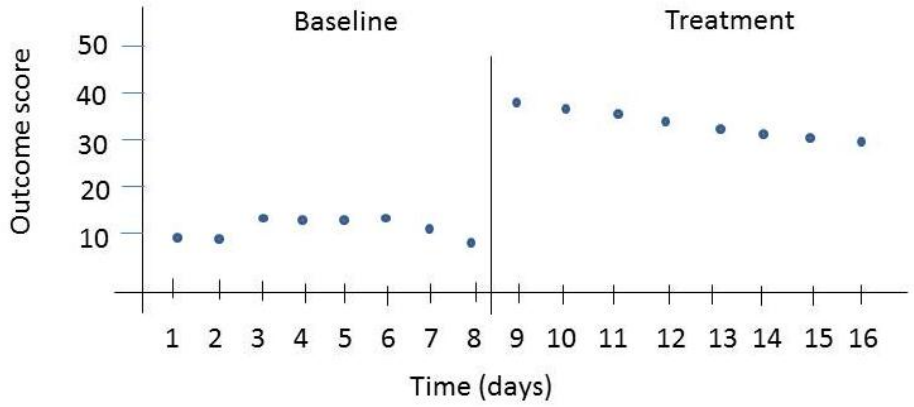

Figure 1.2. Graphical display of the basic AB design.

This basic AB design type is not without criticism. For instance, when using this type of SSED it is difficult to attribute a change in the data to the treatment and not to some other event which could have occurred at the same time (Shadish, Cook, & Campbell, 2002). This limitation can be addressed by utilizing more complex SSED studies, such as multiple-baseline designs, reversal designs, and alternating designs (Barlow et al., 2009). Shadish and Sullivan (2011) conducted a systematic review of 809 published SSEDs in the field of psychology and educational sciences in 2008 and found that more than a half of the SSEDs are characterized by a multiple-baseline design (54.3%). The reversal and alternating treatment designs are the other two most popular SSEDs (8.2% and 8% respectively). Multiple-baseline designs, reversal designs, and alternating treatment designs involve phase repetition and as a result handle major threats towards internal validity including for instance history and maturation (Shadish, et al., 2002).

1.1.2.2 Complex single-subject experimental designs

1.1.2.2.1 Multiple baseline design

A first possible extension of the basic AB phase design, is the multiple-baseline design. In multiple-baseline designs (MBD), an AB phase design (with one baseline phase, A, and one treatment phase, B) is implemented simultaneously to different subjects, behaviors, or settings (Ferron & Scott, 2005; Onghena, 2005; Onghena & Edgington, 2005). The introduction of the treatment is staggered across the subjects, behaviors or settings, which imply baseline phases of different lengths. The general form of an MBD across three subjects with 14 measurement occasions is illustrated in Figure 1.3. MBDs are popular amongst SSEDs (Shadish & Sullivan, 2011) thanks to the sequential introduction of the intervention over the cases (or settings or behaviors). It entails the advantage that researchers can more easily disentangle effects of the intervention and effects of some external events, such as a defective measurement instrument, which leads to more internally valid results (Baer, Wolf, & Risley, 1968; Barlow & Hersen, 1984; Kinugasa, Cerin, & Hooper, 2004; Koehler & Levin, 2000). In Figure 1.3, a decrease in outcome score for the three subjects is observed when the treatment is introduced, independent of the moment in time at which the treatment is administered. The subject's baselines serve as a means of control: the SSED researcher investigates whether a change in outcome scores occurs only for the subject at which the intervention is given and not for the other subjects. Therefore it is more likely that the treatment causes the change in outcome score and not some external factor. Moreover, because the SSED is repeated to several subjects (or behaviors or settings), the external validity of the effectiveness of a treatment can be examined. If the treatment is found to be effective for a group of subjects (or behaviors or settings), a more external (generalizable) treatment effect estimate can be obtained. It can also be the case that the treatment is not effective across the subjects (or behaviors or settings), which gives a motivation to search for moderator variables. In order to examine external validity and moderator variables, typically more than three subjects are needed.

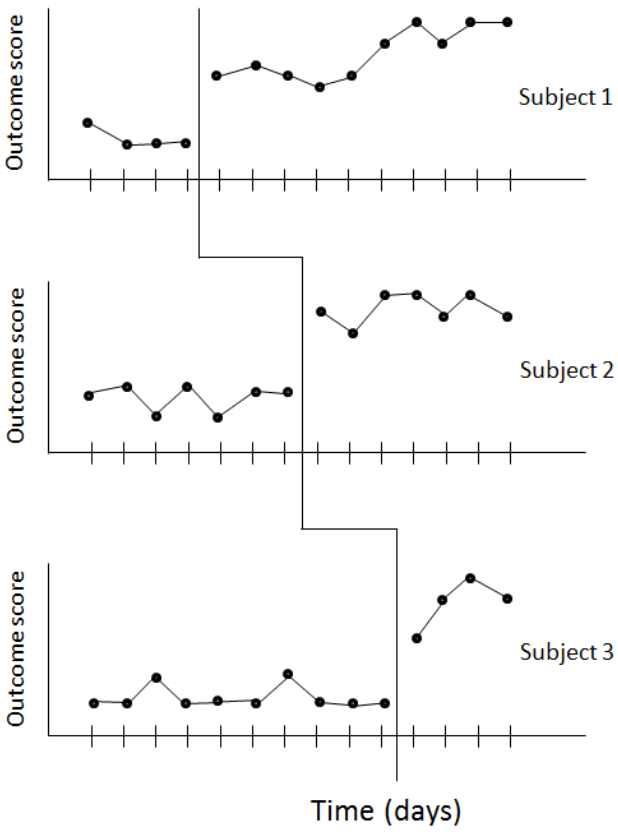

Figure 1.3. Graphical display of the multiple-baseline across three subjects design.

1.1.2.2.2 Reversal designs

The introduction and withdrawal of the treatment is typical for the reversal design (e.g., ABABAB design, see Figure 1.4). In these kinds of designs, there is more than one transition from one phase to another within one subject. If a change in outcome scores after the introduction of the treatment is observed during each AB pair, one can be more confident that a change in outcome scores is due to the treatment and not to some external event effect. The reversal designs provide a high degree of experimental control and are straightforward to implement. But, a drawback is that these designs involve the assumption that the outcome is reversible, which is not always the case. For instance, when the purpose is to learn a new behavior, you cannot unlearn it, and so this type of design gives rise to ethical questions.

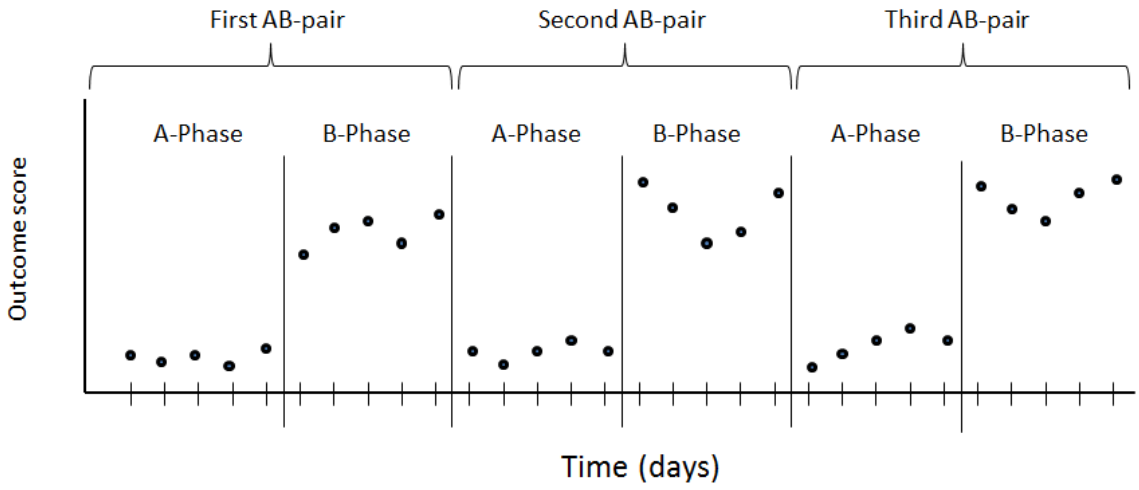

Figure 1.4. Graphical display of the ABABAB reversal design.

A lot of variations of reversal designs are possible in which AB-patterns are replicated (e.g., A-B-A-B-A-B-A-B design), separate treatment variables are evaluated (e.g., A-B-A-C-A design), interaction effects are studied (e.g., A-B-A-B-BC-B-BC design), variations of the same treatment variable are incorporated, (e.g., A-B-A-B-B1-B2-BN design), etc.

1.1.2.2.3 Alternating treatment designs

In many cases however, researchers are not only interested in whether one treatment works but also whether one treatment works better in comparison to another. In an alternating treatment design (ATD), two or more treatments are rapidly alternated (Barlow & Hayes, 1979). In a typical ATD, data collection starts with a baseline phase, but during the treatment phase, two or more treatments are alternated (see Figure 1.5).

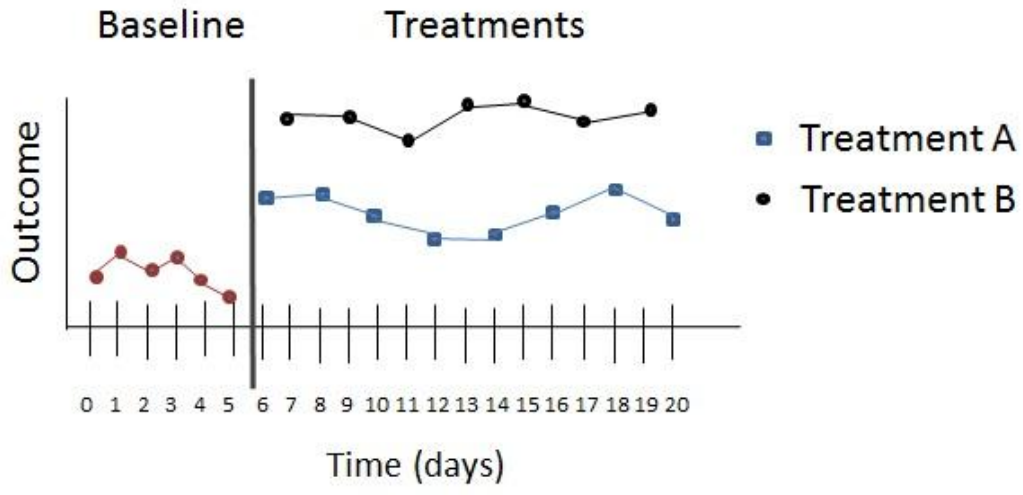

Figure 1.5. Graphical display of the alternating design.

Multiple comparisons of treatments are made in relatively few sessions. Because the dependent variable is exposed to each of the independent variables, carryover effects might occur and it can be questioned whether the treatments per se have an effect. Therefore, a final phase can be included in the design where the selected treatment is implemented alone to ensure that this treatment remains effective. The ATDs entail the advantage that the treatments do not have to be removed, a baseline phase is not needed, and the phases are possibly very short which allows for more quick comparisons. However, this type of design is only appropriate if frequent alternation of the treatments is possible and is therefore a less popular SSED type.

1.2 To Randomize or not to Randomize

An important consideration in designing an SSED is whether or not to incorporate randomization. SSED researchers designing their study could for instance choose to randomly assign measurement occasions to conditions (i.e., for alternating treatment designs) or to randomly choose when the start of a condition occurs (for AB phase designs or reversal designs). As stated by Onghena (2005), the randomization provides statistical control over both known and unknown confounding variables that are time-related (e.g., history and maturation). In this way, randomization can improve the internal validity of an SSED. However, SSEDs are usually nonrandomized experiments, because the random assignment of measurements to conditions or the random start of a condition is practically unfeasible. Therefore caution has to be paid when attributing outcome changes to treatment changes instead of to some external event effect. Advantages of including randomization in the design are described in several textbooks and research articles (e.g., Barlow et al. 2009, Edgington & Onghena, 2007; Kazdin, 2011; Kratochwill & Levin, 1992, 2010). We acknowledge the importance of incorporating randomization in SSEDs to eliminate threats towards internal validity, but we do not limit the work in this dissertation to randomized single case designs.

1.3 The Analysis of a Single-Subject Experimental Design

1.3.1 Visual analysis

Visual analysis of graphed data has been and continues to be the traditional method for evaluating treatment effects in SSED research (Ferron & Jones, 2006; Horner, Swaminathan, Sugar & Stokowski, 2012; Kratochwill et al., 2010), but is by itself less suitable for synthesizing literature in an objective way (Manolov & Solanas, 2013), because it does not provide an effect size measure. Visual analysis methods aim at reaching a judgment about the reliability and consistency of treatment effects by visually examining graphed data. In the *WWC* technical documentation, clear guidelines are reported about which criteria single-case analysts should use to evaluate intervention effects, namely changes in level, variability in outcome scores, trend, the latency of change evident across phases, and whether the changes are consistent with the requirements of the particular design (Kazdin, 2011). When the changes in level, and/or variability are in the desired direction and when they are immediate, readily discernible, and maintained over time, it is concluded that the changes in behavior across phases result from the implemented treatment and are indicative of improvement (Busse et al., 1995). A recent study indicates that visual analyses can lead to consistent results concerning the effectiveness of a treatment only if visual analysts are well trained (Kahn, Chung, Gut shall, Pitts, Kao, & Girolami, 2010). Others have claimed that visual analysis procedures may have Type I error rates that are quite high (Gibson & Ottenbacher, 1988, Greenwood & Matyas, 1990, Matyas & Greenwood, 1990), but the Type I error rate can be controlled in randomized SSEDs by structuring the visual analysis (Ferron & Jones, 2006).

1.3.2 Quantitative analysis

Quantitative analysis methods are still being developed in the domain of SSED research (Kratochwill et al., 2010) and statistical challenges of producing an accepted measure of treatment effect remain (Beretvas & Chung, 2008; Horner, Carr, Halle, McGee, Odom, & Wolery, 2005; Shadish & Rindskopf, 2007).

1.3.2.1 Randomization tests

The use of randomization tests in the area of SSED has been suggested to increase both statistical and internal validity (Bulté & Onghena, 2009; Edgington & Onghena, 2007; Kratochwill & Levin, 2010; Manolov & Solanas, 2009; Onghena, & Edgington, 2005). To conduct a randomization test, researchers have to record all possible random assignments before the start of the SSED study. In alternating treatment designs, measurements are

randomly assigned to phases, or randomly assigned to phases under some restrictions (Onghena & Edgington, 1994). In AB phase or reversal designs the start of a condition is introduced in a random way (Edgington, 1967; Onghena, 1992). In multiple-baseline designs participants can be randomly assigned to baseline lengths (Wampold & Worsham, 1986) or interventions start points can be randomly chosen for each participant subject to some constraints (Koehler & Levin, 1998). One of the possible assignments is chosen and this forms the actual SSED. Then the researcher chooses an appropriate test statistic (e.g., difference in mean outcome between treatment and baseline conditions), collects the data, and calculates the test statistic based on the collected data. Once this is accomplished, the test statistic is calculated for each of the possible alternative random assignments that were recorded at the beginning of the experiment using the collected SSED data. All the test statistic values are sorted and based on this, the statistical significance of the SSED test statistic can be calculated by looking where the obtained test statistic falls within the distribution of possible test statistic values. The *p*-value of the randomized SSED is calculated as the proportion of possible test statistic values that is as extreme as or even more extreme than the value of the test statistic based on the SSED. The use of randomization tests in the context of SSEDs is rather limited because random assignment is not always feasible. Another drawback of these analyses is that the magnitude of a treatment effect cannot be estimated, but one can only decide if the treatment was effective. This can be dealt with by calculating effect size estimates.

1.3.2.2 Effect sizes

In the past, single-case analysts have relied on parametric and nonparametric effect sizes (Maggin, Swami Nathan, Rogers, O'Keeffe, Sugar, & Horner, 2011; Mastropieri & Scruggs, 1985; Methe, Kilgus, Nieman, & Riley-Tillman, 2012). Nonparametric effect sizes such as percentage of non-overlapping data, percentage of all non-overlapping data, or percentage exceeding the median are not affected by distributional assumptions but have been criticized for the inability to (1) account for data trends, (2) discriminate between large treatment effects due to ceiling effects (Wolery, Busick, Reichow, & Barton, 2010), or (3) produce a known sampling distribution (Lenz, 2013; Parker & Vannest, 2008). Parametric effect sizes deal with these critiques and over the last years, several parametric effect sizes have been proposed to enhance the analysis of SSED data including regression estimates (e.g., Maggin, et al., 2011; Parker, Vannest, & Davis, 2011). Amongst others, these methods allow

modeling trends, including predictors, and modeling autocorrelation (Shadish, Rindskopf, Hedges, & Sullivan, 2012).

1.4 The Multilevel Modeling of Single-Subject Experimental Design Data

In the past, little attention is given to the synthesis of SSED results, partly because the fact that the literature about meta-analysis has focused on combining the results of group-comparison studies (Kratochwill et al., 2010). In contrast to SSED studies, in group-comparison studies, there is widespread agreement about how these effect sizes should be expressed, what the statistical properties of the estimators are (e.g., distribution theory, conditional variance), and how to translate from one measure (e.g., a correlation) to another (e.g., Hedges' *g*). However, individual client responses are lost in the group averaging process and important findings are obscured. Inferences about causes of changes (when they can be made) are made at the level of the group, which neglect effects of the intervention on any individual subject. This severely limits the applicability of results to specific clients (Barlow & Hersen, 1984). Group-comparison methods generally involve only one (posttest only) or two (pretest-posttest) measurements of subject response. Important information on the dynamic nature of subject response to treatment is thereby missed.

Therefore, during the last decades, there is a growing interest in synthesizing SSED data across subjects and across studies. A primary search of published meta-analyses of SSED studies using the social sciences citation index within the Web of Sciences using the keywords 'single-case' or 'single-subject' or 'multiple-baseline' in combination with 'meta-analysis' resulted in 2,242 results. Especially during 2012 and 2013, a lot of meta-analyses of SSEDs are reported, see Figure 1.6.

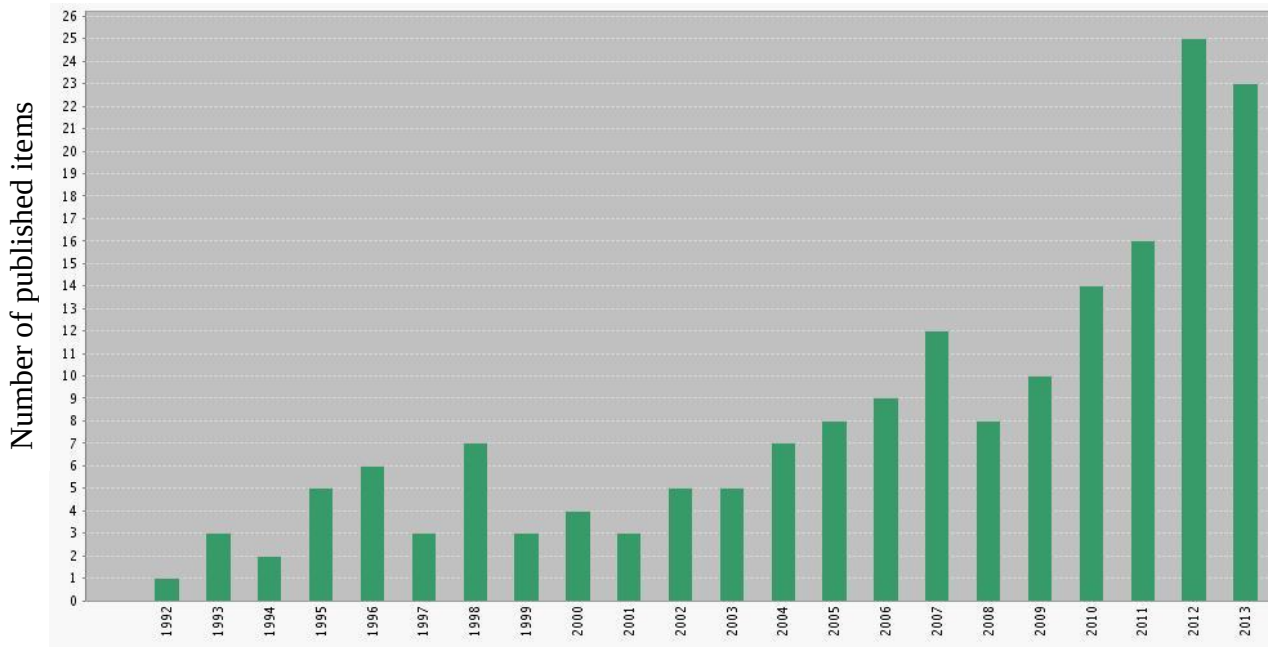

Figure 1.6. Graphical display showing the increase in the number of published items for the keywords "single-case" or "single-subject" or "multiple-baseline" in combination with "meta-analysis".

As the number of published syntheses of SSEDs is increasing, there is a need to optimize the statistical techniques to quantify the research findings in an objective way. When using data from multiple subjects and multiple studies, a three-level structure becomes visible: measurement occasions are nested within subjects and subjects in turn are nested within studies. In SSED studies, data are commonly graphically presented, which allows retrieving the raw data from the primary studies using a statistical software program (e.g., Ungraph, Biosoft, 2004; Shadish, et al., 2009). Afterwards the raw data can be synthesized using a multilevel model which is an extension of the regression approach (Nugent, 1996; Nagler, Rindskopf, & Shadish, 2008; Rindskopf & Ferron, in press; Shadish & Rindskopf, 2007; Shadish, Kyse, & Rindskopf, 2013; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). If the SSED data are not graphically presented in the primary studies, a multilevel analysis can still be conducted based on effect sizes instead of raw data. If effect sizes are combined instead of raw data, we will use the label 'multilevel meta-analysis' instead of 'multilevel analysis' in the remaining of this dissertation.

The multilevel modeling method based on the regression approach is the most flexible approach given its ability to model complexities such as autocorrelation, predictors at the different levels (e.g., age, gender, SES, school type, study quality), heterogeneous within-subject, between-subject and between-study (co)variance, and it allows estimating average treatment effects across studies in addition to subject-specific and study-specific treatment effects (Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a, 2003b, 2008).

By conducting a multilevel analysis, important research questions can be addressed (which cannot be answered by single-level analysis of SSED data) such as: (1) What is the magnitude of the average treatment effect across cases and across studies? (2) What is the magnitude and direction of the case-specific intervention effect? (3) How much does the treatment effect vary within cases, across cases and/or across studies? and (4) Does a (case and/or study level) predictor influence the treatment's effect? The two-level model (Ferron, Bell, Hess, Rendina-Giobioff, & Hibard, 2009; Ferron, Farmer, and Owens, 2010; Van den Noortgate & Onghena, 2003a) and the three-level model (Owens & Ferron, 2010; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013a) have been validated in previous research using extensive simulation studies. Extensions to the three-level model have been proposed, such as the modeling of non-linear trajectories during treatment phase (Beretvas, Hembry, Van den Noortgate, & Ferron, 2013), the modeling of autocorrelation (Baek & Ferron, 2013), standardizing SSED data (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013b), dealing with external event effects (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013c), estimating treatment effect estimates across different types of SSEDs (i.e., multiple-baseline designs, ABAB reversal designs, and alternating treatment designs; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014c), modeling count data as outcome scores (Beretvas & Chu, 2013; Shadish et al., 2013; Shadish & Rindskopf, 2007), etc.

Ignoring the multilayered nature can have a substantial impact on the conclusions of a multilevel analysis (Hox, 2002; Van den Noortgate, Opdenakker, & Onghena, 2005) as standard error estimates will be too small resulting in an inflated number of Type I errors when used in statistical tests (i.e., the statistical test indicates a treatment effect, whereas in reality there is no). Therefore it is important to take the hierarchical structure into account.

1.5 Research Objectives and Structure of this Dissertation

1.5.1 Research objectives

With this doctoral dissertation, we want to contribute to the development of the methodology for combining the results of SSED studies. We suggest, examine, and further extend the multilevel modeling approach to quantitatively integrate SSED data across subjects and across studies. Multilevel modeling of SSEDs allows for a quantitative summary of a large body of literature, which results in externally valid results, more accurate estimates, and valuable information that can inform policy and can improve practice. The intent of this dissertation is twofold. On the one hand, we empirically validate the basic three-level model

and several extensions to it using large simulation studies and giving empirical illustrations. On the other hand, we want to inform applied SSED researchers about the value of multilevel modeling of SSEDs and how to use these models. As a consequence, this dissertation is informative for methodologists, research analysts and synthesists, but also for applied SSED researchers.

1.5.2 Structure of this dissertation

The structure of this dissertation is presented in Figure 1.7.

Chapter 1: General Introduction
Part 1 Three-level modeling: Further developments and methodological issues
Chapter 2: Three-Level Analysis of Unstandardized Single-Subject Data
Chapter 3: Three-Level Analysis of Standardized Single-Subject Data
Chapter 4: Modeling External Event Effects in the Three-Level Modeling of Single-Subject Data
Chapter 5: Misspecification of the Covariance Structure in the Three-Level Modeling of Single-Subject Data
Part 2 Applications
Chapter 6: The Influence of the Design Matrix on Treatment Effect Estimates in the Quantitative Analyses of Single-Subject Data
Chapter 7: From a Single-Level to a Multilevel Analysis of Single-Subject Experimental Data
Chapter 8: Estimating Intervention Effects across Different Types of Single-Subject Designs
Part 3 Discussion, conclusion and future research
Chapter 9: General Discussion
Chapter 10: The future of Multilevel Modeling to Synthesize Single-Subject Experimental Data

Figure 1.7. Overview and structure of this dissertation.

The first part (*Part 1*) of the dissertation is composed of four computer intensive Monte Carlo simulation studies and is especially informative for statisticians, methodologists and SSED research analysts and synthesists. In these simulation studies, we look at the bias of the average treatment effect estimate, which is the difference between the expected effect estimate and the true population effect, at the mean squared error (defined as the mean squared difference between the estimates and the population value), at the standard error estimates, and at the coverage proportion of the 95% confidence intervals for the treatment

effects, which refers to the number of times the 95% confidence interval contains the population value. Furthermore, we examine the power, an important practical consideration when determining the conditions under which the three-level model can be recommended. We also look at the bias of the point estimates of the variance components. For the simulations we use the infrastructure of the Flemish Supercomputer Center, financed by the Department of Economy, Science and Innovation – Flemish Government and the Hercules Foundation. In *Chapter 2*, we evaluate whether the basic three-level model is appropriate to combine raw, unstandardized SSED data across cases and across studies. *Chapter 3* involves the evaluation of a standardizing method in order to combine SSED data over cases and over studies. Standardizing the raw SSED data is needed because dependent variables in a set of SSED studies are not always measured the same way and on the same scale. For instance, challenging behavior in class in one study is measured on a scale from one to ten, whereas another researcher indicates the challenging behavior on a scale from one to five. Standardization allows immediate comparison and fair interpretations of scores on challenging behavior across different studies. In the third simulation study, we focus on the strength of multiple-baseline designs to disentangle treatment and event effects. *Chapter 4* presents a method to adjust the three-level model for external events and evaluates the appropriateness of the modified model. The last chapter of the first part, *Chapter 5*, examines the robustness of the three-level model. The focus of this simulation study is to evaluate the consequences of a violation of independence of the residuals at level two or level three. A major advantage of the multilevel approach is that covariance between the residuals can be modeled by specifying a specific structure for the variances and covariances at either level. We investigate the influence of covariance misspecification on the treatment effect estimates. The purpose of *Part 1* is evaluating the basic three-level model and several extensions to it to synthesize SSED results.

The second part of this dissertation (*Part 2*) consists of three applied studies. In the first chapter of *Part 2* (*Chapter 6)*, the influence of the design matrix specification on the interpretation of the regression coefficients of interest is discussed. Different design matrices are presented that can be used for the most common SSEDs, namely, the multiple-baseline designs, reversal designs, and alternating treatment designs. The purpose of this article is to guide data analysts interested in analyzing and meta-analyzing SSED data. *Chapter 7* goes one step further and explains the process from single-level analysis to multilevel analysis of SSEDs. We advise readers not familiar with multilevel modeling to first read *Chapter 7* because the basics of multilevel modeling are explained in detail. In addition to the basic

multilevel models, several plausible alternative models are elaborated and empirical illustrations are given. Also, a sensitivity analysis is conducted by investigating to what extent the estimated treatment effect is dependent on the modeling specifications and the underlying assumptions. By considering a range of plausible models and assumptions, researchers can determine the degree to which the effect estimates and conclusions are sensitive to the specific assumptions made. If the same conclusions are reached across a range of plausible assumptions, confidence in the conclusions can be enhanced. We end *Part 2* with *Chapter 8* in which we illustrate with the aid of an empirical illustration how SSEDs of several types, including AB phase designs, multiple-baseline designs, ABAB reversal designs, and alternating treatment designs can be combined using the three-level meta-analytic model. The univariate and multivariate three-level meta-analytic models are presented and discussed. If the same conclusion is based on a synthesis of results from different types of SSED designs, then there is more confidence that the results are due to the intervention and not to some outside experimental factors. Combining data from different designs can enhance the external validity of the synthesis' findings because they are based on more diverse data. If several SSED studies' results are combined, then data from multiple studies including one or multiple cases are used thereby providing more information and resulting in more precise treatment effect estimates (i.e., smaller standard errors and narrower confidence intervals).

In the first chapter of the third part (*Part 3*), *Chapter 9*, we give an overview of the main findings and we highlight some strengths of this dissertation, but simultaneously acknowledge that there are a number of limitations. We briefly elaborate implications for applied single-case researchers, research synthesists (meta-analysts) and methodologists, and end *Chapter 9* with a global conclusion. We end this dissertation with *Chapter 10* in which we discuss the future of multilevel modeling to summarize SSED data. In this chapter, we aim to make the reader aware that there is still a lot of work to be accomplished to optimize the multilevel modeling framework, to further extend the multilevel model and to deal with issues highlighted in previous chapters. This chapter can be considered as the beginning of a new research proposal and we hope to stimulate and encourage methodologists, SSED data synthesists, and applied SSED researchers to further study SSEDs and multilevel modeling. In this dissertation we only discovered the top of a huge iceberg.

PART 1|

THREE-LEVEL MODELING: FURTHER DEVELOPMENTS AND METHODOLOGICAL ISSUES

Chapter 2|

Three-Level Analysis of Unstandardized Single-Case Data[1]

Abstract

One approach for combining single-case data within and across studies is multilevel modeling. Although the multilevel approach and its flexibility are appealing, there is much about single-case experimental data and design that is not fully understood. In this article we want to inform research synthesists under which realistic conditions the basic three-level model works to synthesize single-case data. We use an extensive Monte Carlo simulation study to explore the appropriateness of the multilevel modeling inferences. Therefore we choose to vary the value of the immediate treatment effect and the treatment effect on a time trend, the number of studies, the number of cases, the number of measurements per case and the between-case and between-study variance. The simulation study shows that the three-level approach results in unbiased estimates of both kinds of treatment effects. Further, in order to have a reasonable power for testing the treatment effects (.80 or higher), we recommend researchers to use strict inclusion criteria, resulting in a homogeneous set of studies, and to involve a minimum of 30 studies in their three-level analysis of single-case results. The number of measurements and cases is less of importance.

Keywords: single-case study, three-level multilevel analysis, Monte Carlo simulation study

[1] This chapter has been published as Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S.N., & Van den Noortgate, W. (2013a). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education, 82*, 1-21. doi: 10.1080/00220973.2012.745470

2.1 Introduction

2.1.1 Single-case experimental design

A single-case experimental study is "…a designed experiment in which one case is observed repeatedly during a certain period under different levels ('treatments') of at least one independent variable." (Onghena & Edgington, 2005). In the simplest design, an interrupted time series design, a participant or case is repeatedly observed under a baseline condition and a condition during or after a specific treatment. Results are typically graphically displayed, as in Figure 2.1.

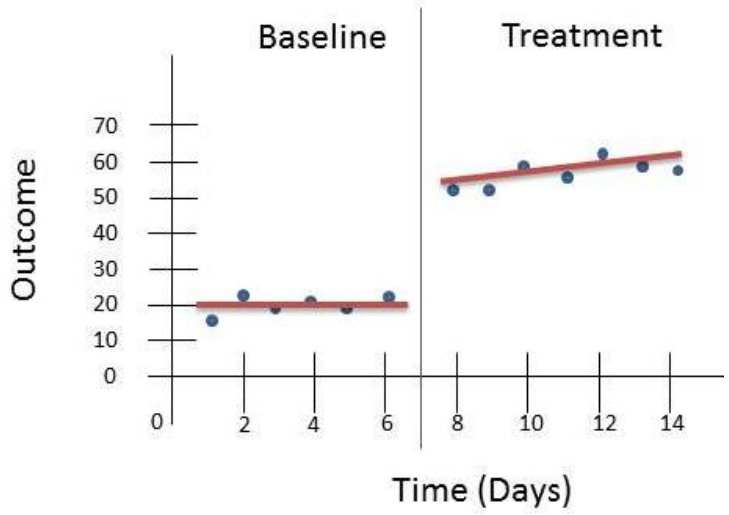

Figure 2.1. Graphical display of the basic single-case experimental design.

The major purpose of this design is to evaluate the effect of the condition on a dependent variable. In the literature, single-case studies have been given a variety of different names, including single-case, N = 1, small-n, intra-subject, single-subject experimental design, interrupted time-series design, among others.

Nowadays there is a growing interest in single-case designs because they have several advantages in comparison to other designs like group comparison designs. This type of research allows researchers to focus on the case-specific treatment effect, which have a tendency to be lost in group comparison designs where the focus is on the average treatment effect across individuals (Barlow & Hersen, 1984). In addition, because the case is measured at various points in time, this type of design also allows researchers to investigate how the treatment effect will change over time. Due to the fact that only one case is needed, researchers are able to study populations that have a low prevalence rate (e.g., children with special needs). Furthermore, because data are collected at multiple points, the researcher can optimize the treatment during the experiment, based on preliminary results. Finally, these

designs reduce the gap between research and practice by allowing practitioners to implement research in their current settings (Morgan & Morgan, 2001).

Although single-case studies are growing in popularity and are valued, they have their limitations. The validity of inferences from basic single-case experimental designs (see Figure 2.1) can be questioned, because a shift in the time series may be the results of something other than the treatment (e.g., an event that happened to occur around the time of the treatment; Shadish et al., 2002). In an effort to reduce the plausibility of alternative explanations for shifts in time series data, single-case researchers often turn to more complex interrupted time-series designs, such as the reversal design and the multiple-baseline design (see Figure 2.2). The reversal design increases the number of phases by withdrawing and reintroducing a treatment, whereas the multiple-baseline design includes interrupted time-series data from multiple participants (or behaviors or settings), where for each participant the treatment begins at a different point in time. The result is that the baselines for the multiple participants are of different lengths. If in a reversal design the performance of the case returns to the baseline level if the treatment is stopped, or if in a multiple-baseline design the change in each time series closely follows the treatment start time, it is not likely that these changes are due to something other than the treatment (Shadish et al., 2002).

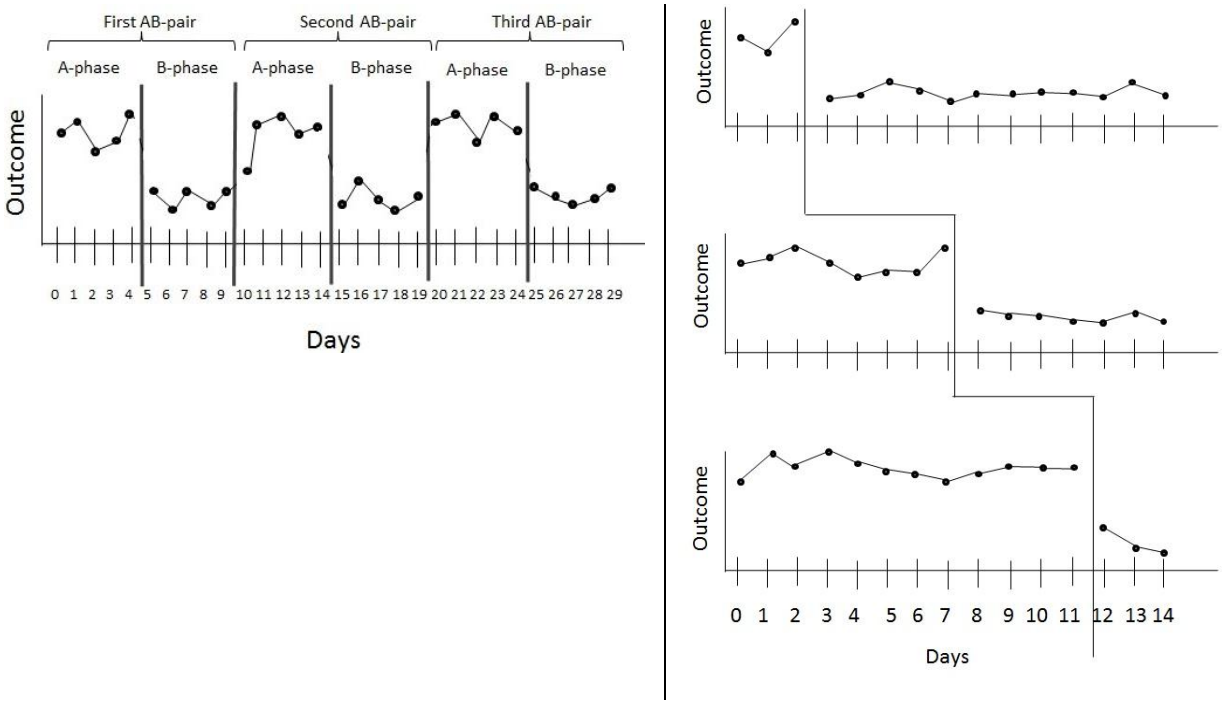

Figure 2.2. Graphical display of the reversal design (left) and the multiple-baseline design (right).

Another limitation of single-case designs is the difficulty of generalizing their results to other cases, because of the small number of cases that are investigated (Kennedy, 1979). To enhance generalizability, researchers replicate across cases, either within studies, such as in a multiple-baseline design, or across studies. As more replications emerge and the evidence base accumulates so does the need for statistical methods designed to synthesize single-case experimental design studies' results and to explore sources of systematic variability in single-cases results by employing moderator analyses.

One approach for combining single-case data within and across studies is multilevel modeling. In a single-case design, the same case is measured repeatedly within and across different conditions or phases (e.g. usually a baseline condition and a treatment condition). Therefore, there are multiple measurements within a case which makes it possible to identify trends in the different conditions. If we than pool together the observations from all cases within one multiple-baseline study, we deal with a two-level structure; measurement occasions at the first level are grouped within cases at the second level. An advantage of using the multilevel approach to synthesize data from multiple cases is that all the measurements within a case are taken into account instead of focusing on the average measure. Therefore the multilevel approach can handle autocorrelation, which means that measurements closer in time are more related to each other than measurements further in time, time trends (e.g. linear or non-linear trends) within each phase of the design and heterogeneous variances (within cases, across cases and across studies). Moreover, we can assess the variation both in the immediate treatment effect and the time trends across cases and across studies. This multilevel approach to single-case data has already been proposed and used in several studies (Ferron et al., 2009; Ferron et al., 2010; Nugent, 1996; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a, 2003b, 2008). To explore the appropriateness of three-level synthesis of single-case data, an initial Monte Carlo study was conducted focusing on the most basic interrupted time series model, one in which there were no trends in either phase (Owens & Ferron, 2012). These studies demonstrate that the fixed effect estimates are unbiased in correctly specified mixed linear models under relatively general conditions when restricted maximum likelihood estimation is used (Kackar & Harville, 1984). In this research we present a more extensive three-level simulation study examining a wider range of conditions than Owens and Ferron (2012) and include trends in both baseline and treatment phases. We also expect unbiased estimates of the treatment effects. Furthermore, we examine the power, an important practical consideration when determining the conditions under which the three-level model can be recommended, which was not examined by Owens and Ferron

(2012). The purpose is to inform applied researchers about the minimal requirements to use three-level modeling for combining the results of single-case data.

In the following sections we present successively the three-level model, the Monte Carlo simulation study and the results from this study, to end with a discussion.

2.1.2 Multilevel analysis of single-case experimental designs

Multilevel structures are ubiquitous in various research areas, for instance in social and behavioral sciences. In educational effectiveness research, a multistage sampling procedure might be used such that in a first stage schools are sampled from a population of schools, next classes from the selected schools and in a third step, students from the selected classes. The students included in the study therefore can be grouped according to the classes and schools they belong to. This structure can induce dependence in the data: students from the same class and the same school are in general more alike than students from different classes and different schools. Therefore, the class and the school membership have to be taken into account when performing statistical analyses. Multilevel models were developed to deal with such grouped data (Raudenbush & Bryk, 2002).

Single-case data from multiple cases also have a hierarchical structure: measurements are grouped in cases. If we have several single-case studies, with more than one case in some studies, three hierarchical levels can be distinguished: measurements are grouped in cases and cases are in turn grouped in studies. The hierarchical structure is illustrated in Figure 2.3.

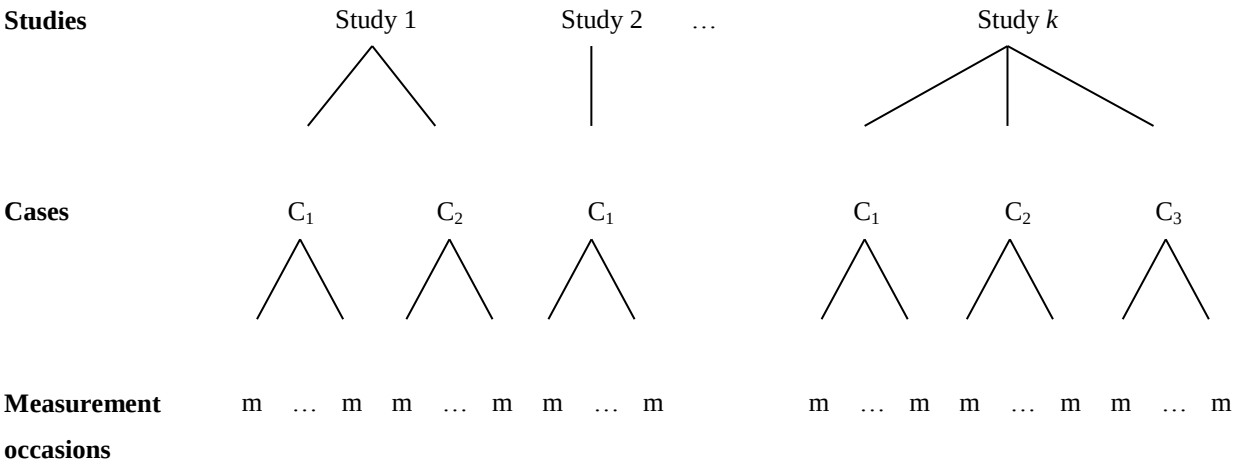

Figure 2.3. The three-level hierarchical structure for single-case experimental design studies.

A multilevel model that can be used to combine single-case data is an extension of the model of Center et al. (1985-1986). More specifically, Van den Noortgate and Onghena (2003a, 2008) suggest to use a hierarchical model in which individual measurements are regressed on a time indicator, *T*, which is centered around the first observation of the

treatment phase, *D*, a dummy variable for the treatment phase, and an interaction term of these variables:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}T_{ijk}D_{ijk} + e_{ijk} \text{ with } e_{ijk} \sim N(0, \sigma_e^2), \quad (2.1)$$

and *i* standing for the measurement occasion (*i* = 0, 1,…*I*), *j* for the case (*j* = 0, 1,…*J*) and *k* for the study (*k* = 0,1,…*K*). The equation shows that in the baseline phase the expected score for the j^{th} case in study *k*, this is $\hat{Y}_{ijk}$, equals $\beta_{0jk} + \beta_{1jk}T_{ijk}$, while it is $(\beta_{0jk} + \beta_{2jk}) + (\beta_{1jk} + \beta_{3jk})T_{ijk}$ in the treatment phase (see Figure 2.4).

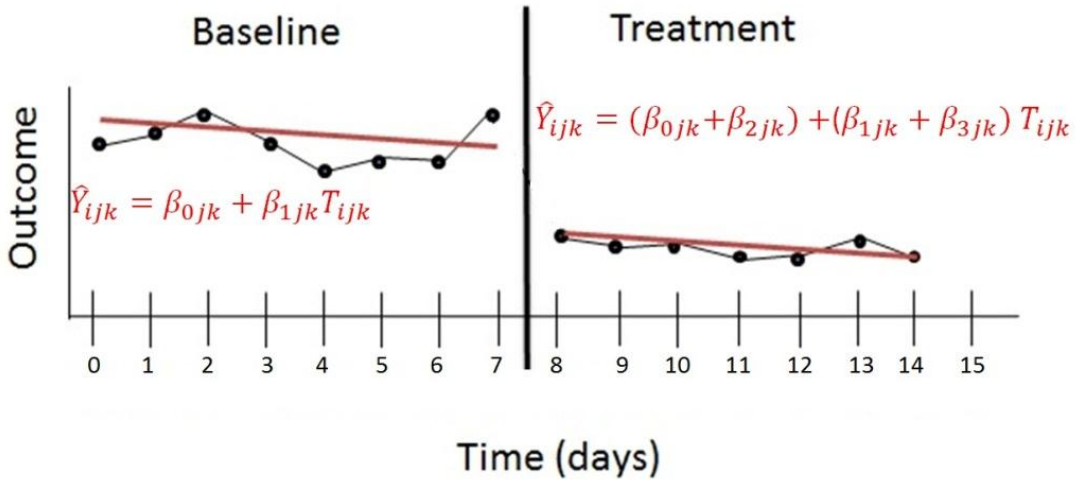

Figure 2.4. Regression model to analyze data from single-case AB phase design.

β_{0jk} therefore indicates the expected baseline level at the start of the treatment phase (when *T* = 0), and β_{1jk} the linear time trend in the baseline scores. The coefficient β_{2jk} can then be interpreted as the immediate effect of the treatment on the outcome, whereas β_{3jk} gives an indication of the effect of the treatment on the time trend.

At the second level of the model, the variation over cases is described using four equations:

$$\begin{cases} \beta_{0jk} = \theta_{00k} + u_{0jk} \\ \beta_{1jk} = \theta_{10k} + u_{1jk} \\ \beta_{2jk} = \theta_{20k} + u_{2jk} \\ \beta_{3jk} = \theta_{30k} + u_{3jk} \end{cases} \text{ with } \begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{2jk} \\ u_{3jk} \end{bmatrix} \sim N(0, \Sigma_u) \quad (2.2)$$

The first equation indicates that the baseline performance for case *j* from study *k* equals an average baseline performance for study *k*, plus a random deviation from this mean; the subsequent equations describe the variation over cases from the same study of the time effect

in the baseline condition, the immediate treatment effect, and the treatment effect on the linear trend, respectively.

At the third level, the variation of the study-specific regression coefficients from the second level equations is described:

$$\begin{cases} \theta_{00k} = \gamma_{000} + v_{00k} \\ \theta_{10k} = \gamma_{100} + v_{10k} \\ \theta_{20k} = \gamma_{200} + v_{20k} \\ \theta_{30k} = \gamma_{300} + v_{30k} \end{cases} \text{ with } \begin{bmatrix} v_{00k} \\ v_{10k} \\ v_{20k} \\ v_{30k} \end{bmatrix} \sim N(0, \Sigma_v) \tag{2.3}$$

Residuals at all three levels are assumed to be multivariate normally distributed.

It is often the case that the studies use different dependent variables. The three-level model can easily be extended by including characteristics of the dependent variable as covariates. This permits a three-level synthesis of single-cases across a set of varying dependent variables, investigating whether the size of the treatment effects depends on the kind of dependent variable. Another way to deal with data measured on different scales is to standardize the data per case by dividing them by the estimated root mean squared error that is found when using the Center et al. (1985-1986) regression model on the data for that case. This method was proposed by Van den Noortgate and Onghena (2003b, 2008).

Parameters of interest are typically primarily the γ's, in multilevel literature called fixed effects, in this case referring to the mean regression coefficients, as well as the (co)variation in the regression coefficients over cases or studies, in multilevel literature called the variance components. This multilevel approach makes it possible to separate sampling variation, between-case variation and between-study variation, and to estimate the treatment effects.

In the current study we are especially interested in the conditions under which the three-level model works acceptably well, which leads us to evaluate the following for the fixed effects: the bias, the mean squared error, the standard error estimates, the coverage proportion for confidence interval estimates, and the power. In addition, we will evaluate bias in the variance component estimates.

2.2 Simulation Study

In order to evaluate in a systematic way the three-level modeling approach, we simulated a number of studies using a multiple-baseline across participants design, based on Equations 2.1 through 2.3. The restricted maximum likelihood procedure in SAS PROC MIXED was used to estimate the three-level model parameters (Littell, Milliken, Stroup,

Wolfinger, & Schabenberger, 2006). The Satterthwaite approach to estimate the degrees of freedom method was used (Satterthwaite, 1941) because this method provides accurate confidence intervals for estimates of the average treatment effect for two level-analyses of multiple-baseline data (Ferron et al., 2009). The procedure used in SAS is a generalization of the Satterthwaite methods described by Giesbrecht and Burns (1985), McLean, Sanders, and Stroup (1988) and Fai and Cornelius (1996). The degrees of freedom are estimated as a function of the covariance matrix of the fixed effects, which is approximated using the covariance matrix of Y along with the design matrix of the fixed effects. Furthermore, the following seven design conditions were varied.

The number of simulated participants per study is equal to 3, 4 or 7 (J = 3, 4 or 7). These numbers were selected based on the recommendation that multiple-baseline studies have at least 3 (Barlow & Hersen, 1984) or 4 baselines (Kazdin & Kopel, 1975), on a survey of multiple-baseline studies (Ferron et al., 2010) which showed studies having from 3 to 10 baselines with a median of 4, on the survey of single-case studies of Shadish and Sullivan (2011), where the number of cases per study ranged from 1 to 13 with median 3, and on a review of Farmer, Owens, Ferron and Allsopp (2010), where 93% of the average number of cases per study fell at or below 7.

The simulated series lengths consist of 10, 20, or 40 measurement occasions (I = 10, 20 or 40). These values were selected based on multiple considerations. A survey of multiple-baseline studies (Ferron et al., 2010) found average series lengths that ranged from 7 to 58 with a median of 24, and a meta-analysis of 85 single-case studies (Swanson & Sachse-Lee, 2000) found that 25 studies had fewer than 11 treatment sessions, 37 studies had between 11 and 29 treatment sessions, and 23 studies had more than 29 treatment sessions. In the survey of Shadish and Sullivan (2011), the number of data points per case ranged from 2 to 160, with median and mode equal to 20, and 90.6% of the cases having 49 or fewer data points. Furthermore, we choose to simulate a number of studies using a multiple-baseline across participants design, therefore we stagger the time of the treatment across cases within studies. The moment at which the treatment starts differs according to the number of cases (J) and the number of measurements (I) within cases (see Table 2.1). For instance, when there are four cases (J = 4) and the number of measurements equals 20 (I = 20), then for the second case the treatment starts on the tenth measurement occasion and lasts until the twentieth measurement occasion.

Table 2.1

The Number of the Measurement Occasion at which the Treatment Started

J	*Case*	*I* = 10	*I* = 20	*I* = 40
3	1	4	7	11
	2	6	11	21
	3	8	15	31
4	1	4	7	11
	2	5	10	18
	3	7	12	24
	4	8	15	31
7	1	4	7	11
	2	5	9	15
	3	5	9	15
	4	6	11	21
	5	7	13	27
	6	7	13	27
	7	8	15	31

The number of simulated studies is 10 or 30 (K = 10 or 30). A review of social science single-case meta-analysis (Farmer et al., 2010) showed that the number of studies included in a meta-analysis ranged from 3 to 117, with 60% of the meta-analysis including less than 30 studies. We chose to include only lower limits for the number of studies (K = 10 or 30) to test if the model works appropriate in these conditions.

The within person variance, is set to 1.0. In this way, the values chosen for the regression coefficients can also be regarded as the expected coefficients standardized by dividing by the residual within case standard deviation.

The immediate effect of the treatment on the outcome, γ_{200}, was generated to have values of 0 (no effect) or 2. In re-analyses of meta-analyses (Alen, Grietens, & Van den Noorgate, 2009; Denis, Van den Noortgate, & Maes, 2011; Kokina & Kern, 2010; Shogren, Fagella-Luby, Bae, & Wehmeyer, 2004; Wang, Cui, & Parrila, 2011), we found similar values for standardized regression coefficients. The effect of the treatment on the trend, defined by γ_{300}, was varied to have values equal to 0 (representing no effect) or 0.2, based on our analyses of real data sets (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011). The regression coefficients of the baseline γ_{000} and γ_{100} were not varied (and are set at 0), because the focus of the current study is on treatment effects.

The between-case covariance matrix Σ_u, was manipulated to have conditions with relatively small and relatively large amounts of between-case variance. Covariances between regression coefficients were set to zero at the subject and study level. Therefore, Σ_u is a diagonal matrix, $\Sigma_u = diag(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3})$. A review of several re-analyses showed that the variance between participants is sometimes less than the within-person variance (Ferron, et al., 2009; Van den Noortgate & Onghena, 2003b) and sometimes greater than the within-person variance (Van den Noortgate & Onghena, 2008). Therefore we vary the four diagonal elements of Σ_u (the variances in the baseline intercept, the baseline slope, the immediate treatment effect and the treatment effect on the time trend respectively) as follow: $\Sigma_u = \text{diag}(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}) = \text{diag}(2, 0.2, 2, 0.2)$, representing a relatively large amount of between-case variability (compared to the within-person variance of one) and $\Sigma_u = \text{diag}(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}) = \text{diag}(0.5, 0.05, 0.5, 0.05)$ to represent a relatively small amount of between-case variability. Re-analyses of real data sets (Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004) indicated that the variance of the effect of γ_{200} is sometimes much larger than the variance of the effect of γ_{300}. Therefore, we also choose to set $\Sigma_u = \text{diag}(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}) = \text{diag}(8, 0.08, 8, 0.08)$.

Again based on re-analyses of meta-analyses (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011), we have chosen the same sets of values for the four diagonal elements of the between-study variance: $\Sigma_v = \text{diag}(\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}) = \text{diag}(2, 0.2, 2, 0.2)$, or $\Sigma_v = \text{diag}(\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}) = \text{diag}(0.5, 0.05, 0.5, 0.05)$, or $\Sigma_v = \text{diag}(\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}) = \text{diag}(8, 0.08, 8, 0.08)$.

In total we therefore have 3x3x2x2x2x3x3 = 648 combinations. For each combination, we simulated 2,000 datasets, 1,296,000 in total. Each dataset was analyzed using the three-level model used to generate the data (Equations 2.1 - 2.3).

2.3 Results of the Simulation Study

We will present the results in two sections. The first section presents the bias of the point estimate, the mean squared error, the estimation of the standard error, the coverage proportion of the confidence interval, and the power for the estimates of the average effects. The second section presents the bias of the point estimates of the variance components.

2.3.1 Average treatment effects

2.3.1.1 Bias and mean squared error

Figure 2.5 and Figure 2.6 show the distribution of the deviations of the estimated immediate treatment effect from its population value (γ_{200}) and the treatment effect on the time trend from its population value (γ_{300}).

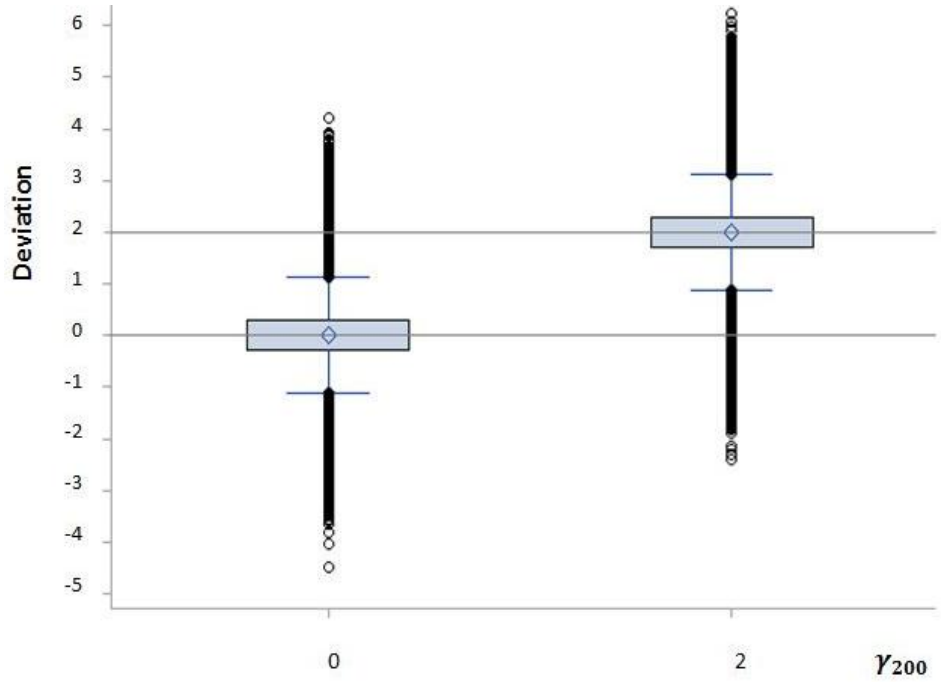

Figure 2.5. Distribution of the deviations of the estimated immediate treatment effect from its populations value (γ_{200}).

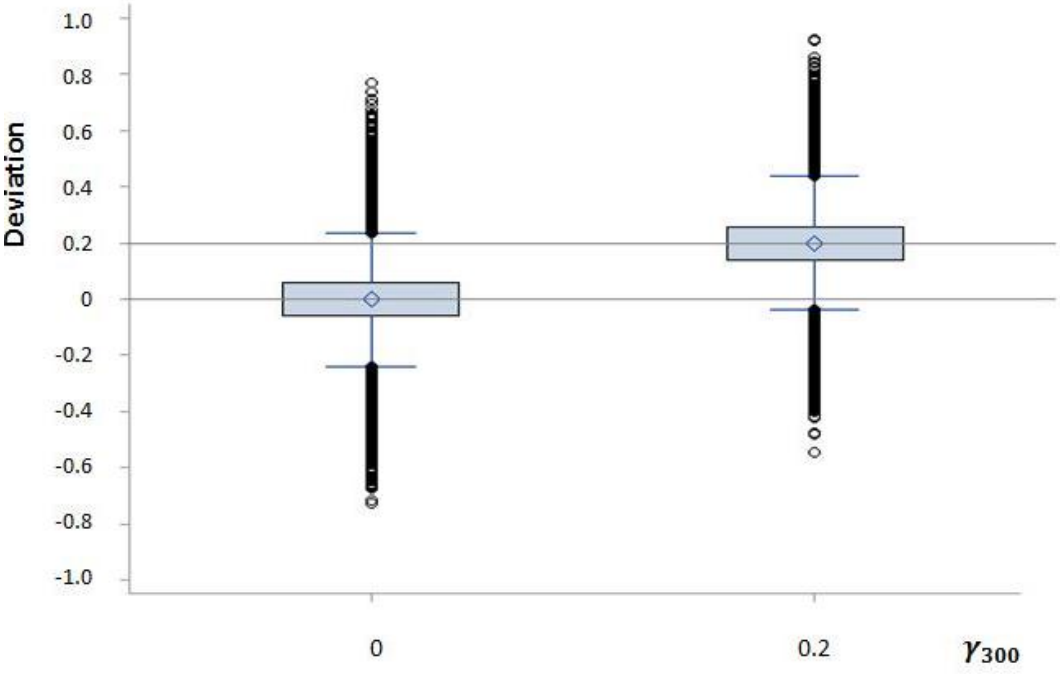

Figure 2.6. Distribution of the deviations of the estimated effect of treatment on the time trend from its populations value (γ_{300}).

We investigated the absolute bias and the relative bias for the estimates of the treatment effects, γ_{200} and γ_{300}. The absolute bias is the difference between the expected effect estimate and the true population effect. The relative bias is the absolute bias divided by the population parameter value. We expected both absolute bias and relative bias to be zero, which is what was found. When γ_{200} and γ_{300} equal 0, the estimated absolute bias was respectively 0.00065 and -0.0000036 and when γ_{200} = 2 and γ_{300} = 0.2, the relative bias was respectively 0.00017 and 0.00060.

The Mean Squared Error (*MSE*) of the average effect estimates, defined as the mean squared difference between the estimates and the population value, gives important information about both bias and variance of the estimates. The smaller the *MSE*, the better the estimate. Table 2.2 provides the *MSE* for the immediate effect of the treatment (γ_{200}) for a subset of the conditions, specifically for those in which γ_{200} = 2 and γ_{300} = 0.2. Similar patterns were seen for the other combinations of fixed effect values (and the full set of results is available from the first author).

Table 2.2

Mean Squared Error of γ_{200} for γ_{200} = 2 and γ_{300} = 0.2 Conditions

			K = 10			*K* = 30		
I	*J*	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8
10	3	0.5	0.05	0.13	0.43	0.02	0.04	0.13
		2	0.08	0.16	0.40	0.03	0.05	0.15
		8	0.16	0.22	0.48	0.06	0.09	0.17
	4	0.5	0.05	0.11	0.37	0.01	0.04	0.13
		2	0.06	0.13	0.43	0.02	0.04	0.13
		8	0.13	0.20	0.46	0.04	0.07	0.15
	7	0.5	0.04	0.11	0.39	0.01	0.04	0.13
		2	0.05	0.13	0.38	0.02	0.04	0.13
		8	0.09	0.16	0.44	0.03	0.06	0.14
20	3	0.5	0.04	0.11	0.35	0.02	0.04	0.13
		2	0.07	0.13	0.43	0.02	0.05	0.14
		8	0.17	0.24	0.47	0.05	0.07	0.17
	4	0.5	0.04	0.11	0.38	0.01	0.04	0.13
		2	0.06	0.13	0.41	0.02	0.04	0.14
		8	0.13	0.18	0.46	0.04	0.06	0.15
	7	0.5	0.03	0.11	0.35	0.01	0.03	0.12
		2	0.04	0.11	0.38	0.01	0.03	0.13
		8	0.09	0.15	0.40	0.03	0.05	0.15
40	3	0.5	0.04	0.10	0.35	0.01	0.04	0.12
		2	0.06	0.14	0.42	0.02	0.04	0.13
		8	0.15	0.22	0.49	0.05	0.08	0.16
	4	0.5	0.03	0.10	0.34	0.01	0.03	0.12
		2	0.05	0.11	0.42	0.02	0.04	0.14
		8	0.12	0.20	0.45	0.04	0.06	0.16
	7	0.5	0.03	0.10	0.34	0.01	0.03	0.12
		2	0.04	0.11	0.39	0.01	0.04	0.12
		8	0.08	0.12	0.43	0.03	0.05	0.14

An important finding is that the *MSE* for the treatment effects, γ_{200} and γ_{300}, became about three times smaller if the number of studies increased from 10 to 30. For instance, if the number of studies was set on 10, the *MSE* equaled 0.13 when *I* = 10, *J* = 3, $\sigma^2_{v_2}$ = 2 and $\sigma^2_{u_2}$ = 0.5, whereas it was 0.04 when the number of studies was set on 30. If the analysis involved only 10 studies instead of 30, the *MSE* was affected by the number of cases. A remarkable finding was that if we only have 10 studies, the *MSE* was reduced if studies consisted of 4 cases rather than 3, but that it was hardly reduced further by increasing the number of cases per study to 7. For instance, if the number of cases was set on 3, the *MSE* for the condition where *K* = 10, *I* = 10, $\sigma^2_{v_2}$ = 2, $\sigma^2_{u_2}$ = 2, equaled 0.16, whereas the *MSE* equaled 0.13 for 4 and 7 cases. Yet, when the analysis involved 30 studies instead of 10, the number of cases hardly

influenced the *MSE* at all. To illustrate, if *K* = 30, *I* = 10, $\sigma^2_{v_2}$ = 2 and $\sigma^2_{u_2}$ = 2, the *MSE* for 3, 4 and 7 cases equaled respectively 0.05, 0.04 and 0.04. This finding is very important in practice. If we have enough studies, the size of the study hardly matters. The *MSE* further is affected by the size of the between-study and the between-case variance (see Table 2.2). The larger the between-case and especially the between-study variance, the larger was the *MSE*. The patterns were very similar for the estimates of γ_{300}, only the *MSE* is about 10 times smaller.

2.3.1.2 Estimates of the standard errors of the average effects

In order to construct confidence intervals around the estimated effects, γ_{200} and γ_{300}, the standard error for these effect estimates can be estimated. By definition, the standard error equals the standard deviation of the sampling distribution of the effect estimator. In this study, we simulated for each condition 2,000 data sets, which resulted in 2,000 estimates of the effects and 2,000 estimates of the corresponding standard error. Because of this relatively large number of estimates, the standard deviation of these effect estimates can be regarded as a good estimate of the standard deviation of the sampling distribution, and can therefore be used as a criterion to evaluate the estimated standard errors.

Figure 2.7 shows that the median of the standard error estimates for γ_{200} is almost equal to the standard deviation of the estimates of the effect for different values of the number of studies (*K*) and the between-study covariance matrix (Σ_v). When the number of studies increased (*K* = 30 versus 10) and the between-study variance decreased ($\sigma^2_{v_2}$= 0.5 versus 8), the standard error decreased (*SE* = 0.20 versus 0.90).

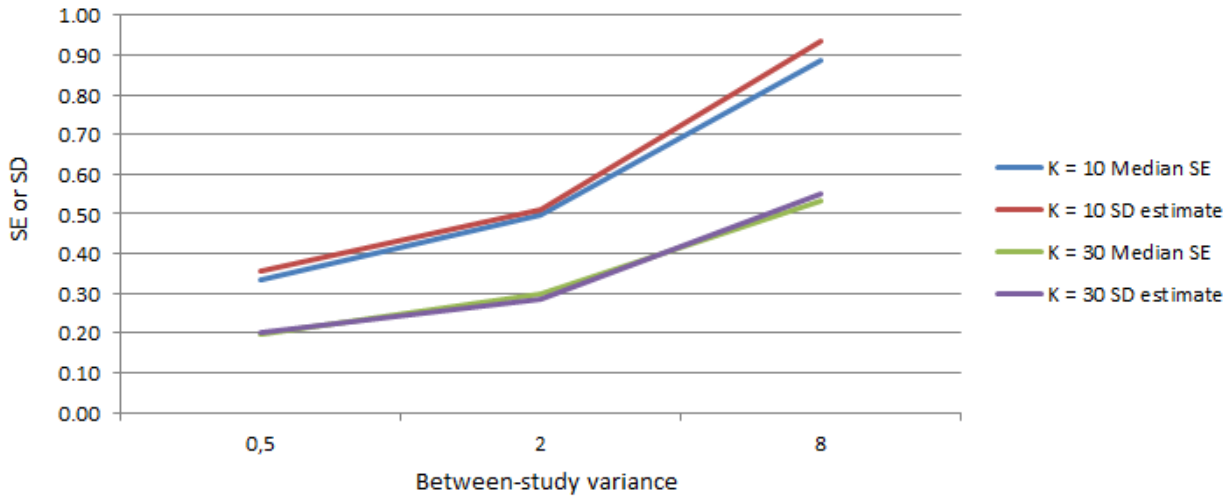

Figure 2.7. Median of standard error compared to standard deviation of the estimates of the effect of γ_{200} , for γ_{200} = 2, γ_{300} = 0.2, J = 4, I = 20 and $\sigma^2_{u_2}$ = 2 conditions.

The patterns were similar for the estimates of the standard error of γ_{300} (see Figure 2.8), but the values of the standard errors were much smaller.

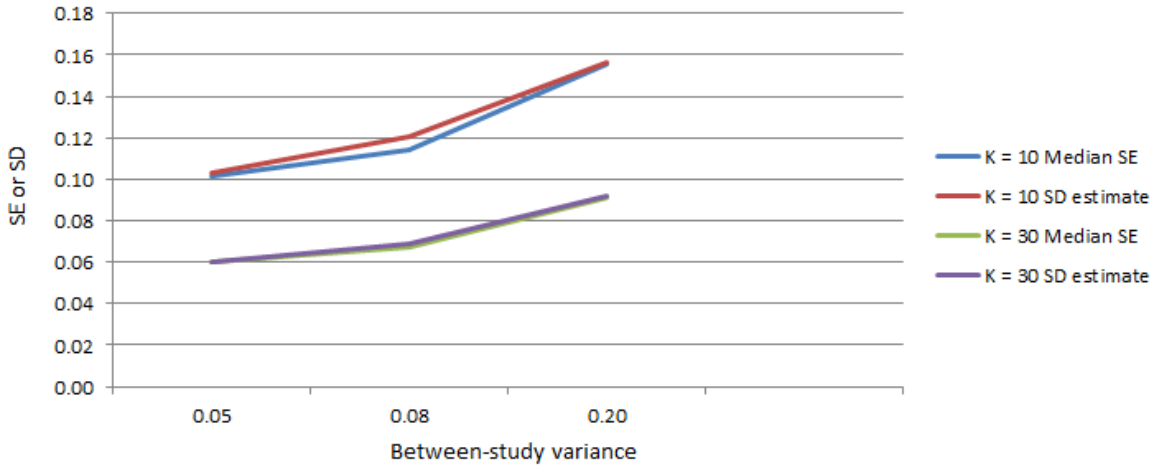

Figure 2.8. Median of standard error compared to standard deviation of the estimates of the effect of γ_{300} , for γ_{200} = 2, γ_{300} = 0.2, J = 4, I = 20 and $\sigma^2_{u_3}$ = 0,2 conditions.

2.3.1.3 Coverage proportion

Another way to evaluate the interval estimates of the estimated effects γ_{200} and γ_{300} and their standard error estimates is to estimate the coverage proportion of the confidence intervals that are calculated using these standard errors and the Satterthwaite estimated degrees of freedom. More specifically, we calculated the proportion of 95% confidence intervals around the effect estimates that contain the population values of γ_{200} and γ_{300}.

The estimated coverage proportion for both γ_{200} and γ_{300} ranged from .93 to .97 with a median of .95 and a standard deviation of 0.0055 and 0.005 respectively. These deviations from the nominal value of .95 can be explained by chance (the standard deviation of the proportions is not larger than the standard error for proportions calculated on 2,000

observations. If the population proportion is equal to .95: *SE* = $\sqrt{(0.95*0.05)/2{,}000}$ = 0.005).

2.3.1.4 Power

The probability of rejecting the null hypothesis when in fact a certain alternative parameter value is true, is called the power of a significance test (Cohen, 1988). When the null hypothesis is true (γ_{200} = 0) and α equals .05, we expect the power (i.e., the Type I error rate) to be equal to .05. We found that the actual Type I error rate was close to .05 in all conditions.

When the null hypothesis is false (γ_{200} = 2) and α equals .05, we want the power as high as possible. A power equal to or higher than .80 is often regarded as an acceptable degree of power (Cohen, 1988). For decisions about the research design it is useful to estimate which conditions are needed to achieve this specific power. In Table 2.3, conditions in which this power level of .80 was reached are marked in bold. When the number of studies equals 30, the power was equal to or larger than .88 for all conditions. This did not apply when only 10 studies were involved. When there were 10 studies included, the power estimates did not reach the threshold of .80 when the between-study variance was large ($\sigma^2_{v_2}$= 80) and number of measurements was large (*I* = 40), independent of the number of cases and the between-case variance. In these conditions the power had values between .39 and .54. When there were 10 studies included, the power reached the threshold of .80 in contexts where the studies are quite homogeneous ($\sigma^2_{v_2} = 0.5$). When there were 10 studies and the between-study variance equaled 2, power sometimes reached the threshold of .80, but tended to be lower in conditions with a large between-case variance ($\sigma^2_{u_2}$= 8). In these conditions the power had values between .41 and .93. So the number of studies and the between-study variance appear to be of particular importance for achieving a power of .80.

The actual Type I error rate for γ_{300} was close to its nominal value of .05. In the contexts where the null hypothesis was false (γ_{300} = 0.2) and the number of studies was 10, the power estimates had values between .17 and .66. This means that the threshold of .80 was not reached in any of the conditions (see Table 2.4). The power estimates lay between .75 and .99 when 30 studies were involved and the number of measurements was 20 or 40 in combination with homogeneous studies ($\sigma^2_{v_3}$ = 0.05 or 0.08). This suggests the importance of including enough studies in the multilevel modeling of single-case studies.

Table 2.3

Power of γ_{200} for γ_{200}=2 and γ_{300} = 0.2 Conditions

			J = 3			*J = 4*			*J = 7*		
K	*I*	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8
10	10	0.5	**1.00**	**.93**	**1.00**	**.95**	.47	**1.00**	**.96**	.48	.48
		2	**.99**	**.88**	**1.00**	**.91**	.50	**1.00**	**.94**	.49	.49
		8	**.84**	.69	**.91**	.77	.41	**0.99**	**.88**	.47	.47
	20	0.5	**1.00**	**.94**	.52	**1.00**	**.96**	**1.00**	**.96**	.51	.51
		2	**1.00**	**.92**	.47	**1.00**	**.94**	**1.00**	**.94**	.50	.50
		8	**.85**	.70	.40	**.92**	**.80**	**.99**	**.87**	.46	.46
	40	0.5	**.00**	**.96**	.51	**1.00**	**.97**	.54	**1.00**	**.97**	.50
		2	**1.00**	**.92**	.47	**1.00**	**.94**	.49	**1.00**	**.96**	.51
		8	**.87**	.73	.39	**.94**	.79	.41	**.99**	**.88**	.47
30	10	0.5	**1.00**	**1.00**	**.95**	**1.00**	**1.00**	**.95**	**1.00**	**1.00**	**.96**
		2	**1.00**	**1.00**	**.94**	**1.00**	**1.00**	**.95**	**1.00**	**1.00**	**.95**
		8	**1.00**	**1.00**	**.88**	**1.00**	**1.00**	**.91**	**1.00**	**1.00**	**.95**
	20	0.5	**1.00**	**1.00**	**.96**	**1.00**	**1.00**	**.96**	**1.00**	**1.00**	**.96**
		2	**1.00**	**1.00**	**.93**	**1.00**	**1.00**	**.94**	**1.00**	**1.00**	**.96**
		8	**1.00**	**1.00**	**.89**	**1.00**	**1.00**	**.91**	**1.00**	**1.00**	**.94**
	40	0.5	**1.00**	**1.00**	**.96**	**1.00**	**1.00**	**.96**	**1.00**	**.97**	**.96**
		2	**1.00**	**1.00**	**.95**	**1.00**	**1.00**	**.96**	**1.00**	**.96**	**.95**
		8	**1.00**	**1.00**	**.90**	**1.00**	**1.00**	**.91**	**.99**	**.88**	**.94**

Note. Values ≥ .80 are in boldface.

Table 2.4

Power of γ_{300} for γ_{300}=0.2 and γ_{200} = 2 Conditions

			J = 3			*J = 4*			*J = 7*		
K	*I*	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 0.05	$\sigma^2_{v_2}$ = 0.08	$\sigma^2_{v_2}$ = 0.2	$\sigma^2_{v_2}$ = 0.05	$\sigma^2_{v_2}$ = 0.08	$\sigma^2_{v_2}$ = 0.2	$\sigma^2_{v_2}$ = 0.05	$\sigma^2_{v_2}$ = 0.08	$\sigma^2_{v_2}$ = 0.2
10	10	0.05	.33	.28	.21	.39	.32	.20	.47	.36	.22
		0.08	.30	.25	.17	.36	.30	.20	.44	.36	.21
		0.2	.23	.20	.17	.29	.25	.17	.38	.33	.20
	20	0.05	.53	.40	.21	.58	.41	.23	.63	.45	.24
		0.08	.47	.38	.21	.51	.39	.22	.60	.45	.24
		0.2	.35	.28	.19	.38	.31	.19	.49	.39	.23
	40	0.05	.58	.43	.24	.59	.47	.20	.66	.48	.24
		0.08	.52	.41	.24	.55	.44	.22	.60	.46	.26
		0.2	.36	.30	.20	.41	.36	.20	.50	.41	.21
30	10	0.05	**.81**	.72	.50	**.87**	.79	.53	**.95**	**.86**	.57
		0.08	.75	.69	.46	**.84**	.75	.52	**.93**	**.84**	.56
		0.2	.65	.57	.41	.73	.66	.47	**.88**	**.80**	.54
	20	0.05	**.96**	**.89**	.59	**.98**	**.91**	.61	**.99**	**.94**	.62
		0.08	**.95**	**.87**	.58	**.97**	**.90**	.60	**.99**	**.92**	.63
		0.2	**.83**	.75	.52	**.89**	**.81**	.55	**.95**	**.88**	.58
	40	0.05	**.98**	**.93**	.62	**.99**	**.93**	.62	**.99**	**.95**	.64
		0.08	**.96**	**.89**	.61	**.98**	**.92**	.62	**.99**	**.94**	.65
		0.2	**.87**	**.79**	.55	**.92**	**.84**	.59	**.97**	**.89**	.60

Note. Values ≥ .80 are in boldface.

2.3.2 Variance components

In the three-level analyses, the between-study and between-case variances were estimated for both the effect of the treatment on the intercept and the trend.

In some conditions convergence was not reached and the last variance estimates were unrealistically large. Therefore we deleted the variance estimates larger than 100 (this was the case in 7.10^{-5} % of the datasets). The variance estimates were still positively skewed after deletion of the extreme values (skewness = 2.38) partly due to truncation of negative estimates to zero. Because variance estimates are still expected to be positively skewed, due to truncation of negative estimates to zero, we calculated the median (relative) deviation of the estimates from the population value, rather than the mean (relative) deviation, to evaluate the (relative) bias in the estimates.

Table 2.5 shows that there is negative relative bias in the estimated between-study variance ($\sigma_{v_2}^2$) and the estimated between-case variance ($\sigma_{u_2}^2$) of the immediate effect and that this bias was worse when there are only 10 studies involved, the between-study variance is small ($\sigma_{v_2}^2$ = 0.5), and the between-case variance is large ($\sigma_{u_2}^2$ = 8). In these conditions, the estimated between-study variance had relative bias values between -0.21 and -0.50. Although variance estimates larger than 100 were not included in the calculation, substantially large bias values were obtained when estimating the between-study variance with bias as large as 50% in the condition where K = 10, I = 20, J = 3, $\sigma_{u_2}^2$= 8 and $\sigma_{v_2}^2$= 0.5. This is an extremely high relative bias and can be explained by the positively skewed distribution of the estimated variance components. When estimating the between-case variance of the immediate effect, similar conclusions can be made, but the relative bias was smaller, with a maximum of 10%. The estimate of the between-study and the between-case variance of the effect on the time trend lead to similar conclusions, except that there is a positive instead of negative relative bias.

Table 2.5

Median of Relative Deviation of the Variance Estimates of γ_{200}, for γ_{200} = 2 and γ_{300} = 0.2 Conditions

			$\hat{\sigma}^2_{v_2}$						$\hat{\sigma}^2_{u_2}$					
			K = 10			*K* = 30			*K =10*			*K = 30*		
I	*J*	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 0.2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 0.2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 0.2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 0.2	$\sigma^2_{v_2}$ = 8
10	3	0.5	-0.16	-0.09	-0.08	-0.02	-0.03	-0.03	-0.10	-0.08	-0.07	-0.05	-0.04	-0.04
		2	-0.23	-0.07	-0.10	-0.04	-0.01	-0.03	-0.08	-0.08	-0.05	-0.02	0.00	0.00
		8	-0.42	-0.17	-0.08	-0.22	-0.05	-0.02	-0.09	-0.06	-0.03	-0.03	-0.01	-0.01
	4	0.5	-0.15	-0.08	-0.08	-0.03	-0.02	-0.04	-0.08	-0.04	-0.08	-0.03	-0.03	-0.03
		2	-0.22	-0.06	-0.09	-0.03	-0.02	-0.03	-0.04	-0.06	-0.05	-0.02	-0.01	-0.01
		8	-0.27	-0.12	-0.10	-0.09	-0.07	-0.03	-0.05	-0.04	-0.03	-0.02	-0.00	-0.00
	7	0.5	-0.10	-0.09	-0.07	-0.04	-0.02	-0.02	-0.06	-0.04	-0.06	0.00	-0.01	-0.01
		2	-0.09	-0.10	-0.07	-0.02	-0.01	-0.01	-0.01	-0.01	-0.04	-0.01	-0.01	-0.01
		8	-0.21	-0.14	-0.07	-0.04	-0.02	-0.02	-0.03	-0.02	-0.01	-0.01	-0.01	-0.01
20	3	0.5	-0.11	-0.10	-0.08	-0.05	-0.03	-0.02	-0.07	-0.07	-0.07	-0.03	-0.01	-0.01
		2	-0.17	-0.11	-0.09	-0.03	-0.02	0.00	-0.08	-0.05	-0.04	-0.01	-0.00	-0.00
		8	-0.50	-0.11	-0.08	-0.09	-0.04	-0.02	-0.09	-0.06	-0.04	-0.04	-0.01	-0.01
	4	0.5	-0.12	-0.09	-0.09	-0.04	-0.02	-0.03	-0.03	-0.07	-0.03	-0.01	-0.00	-0.00
		2	-0.16	-0.12	-0.11	-0.07	-0.01	-0.03	-0.07	-0.03	-0.03	-0.00	-0.01	-0.02
		8	-0.27	-0.16	-0.08	-0.11	-0.05	-0.03	-0.06	-0.04	-0.03	-0.02	-0.01	-0.01
	7	0.5	-0.07	-0.08	-0.08	-0.03	-0.03	-0.04	-0.02	-0.02	-0.00	-0.01	-0.01	-0.01
		2	-0.16	-0.06	-0.08	-0.04	-0.03	-0.02	-0.02	-0.02	-0.01	-0.00	-0.00	-0.00
		8	-0.24	-0.12	-0.09	-0.13	-0.04	-0.03	-0.03	-0.01	-0.02	-0.00	-0.00	-0.00
40	3	0.5	-0.13	-0.09	-0.09	-0.03	-0.03	-0.03	-0.08	-0.04	-0.06	-0.03	-0.01	-0.01
		2	-0.16	-0.12	-0.06	-0.05	-0.03	-0.01	-0.04	-0.04	-0.05	-0.03	-0.01	-0.01
		8	-0.49	-0.18	-0.11	-0.10	-0.03	-0.03	-0.09	-0.07	-0.02	-0.03	-0.02	-0.02
	4	0.5	-0.05	-0.11	-0.09	-0.03	-0.02	-0.01	-0.03	-0.04	-0.04	-0.00	-0.01	-0.01
		2	-0.13	-0.09	-0.09	-0.05	-0.03	-0.01	-0.04	-0.02	-0.03	-0.01	-0.00	-0.00
		8	-0.42	-0.09	-0.10	-0.11	-0.05	-0.05	-0.06	-0.04	-0.02	-0.01	-0.01	-0.01
	7	0.5	-0.10	-0.09	-0.07	-0.03	-0.03	-0.02	-0.03	-0.02	-0.01	-0.01	-0.01	-0.01
		2	-0.11	-0.06	-0.08	-0.05	-0.03	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01
		8	-0.28	-0.11	-0.09	-0.11	-0.03	-0.02	-0.02	-0.01	-0.01	-0.01	-0.00	-0.00

2.4 Discussion

2.4.1 General conclusion

The purpose of this simulation study was to examine the recovery of parameter and standard error estimates for the three-level model used with single-case data as proposed by Van den Noortgate and Onghena (2008). The study examined estimation of the fixed effects (i.e., the average immediate effect at the start of the treatment and the average effect on the time trend) and the variance components (i.e., the between-cases within-study variance and the between-study variance in the immediate effect and in the treatment effect on the time trend).

The results indicated that regardless of the condition, the fixed effect estimates are unbiased. This finding was theoretically expected (Kackar & Harville, 1984) and is consistent with previous research regarding estimation of fixed effects in two level models (Ferron et al., 2009; Raudenbush & Bryk, 2002). Further, the medians of the standard errors of γ_{200} and γ_{300} are almost equal to the standard deviations of the estimates of the effects for different values of the number of studies and the between-study variance. The standard error decreases when the number of studies increases and when the between-study variance is more homogeneous. In all conditions the coverage proportion has values between 93% and 97% for both the estimation of the effect on the intercept and the effect on the trend. These findings correspond to previous research concerning the two-level analysis of single-case data (Ferron et al., 2009) and concerning multilevel synthesis of more traditional longitudinal designs (Fouladi & Shieh, 2004; Gomez, Schaalje, & Fellingham, 2005; Kowalchuk, Keselman, Algina, & Wolfinger 1997). An important finding is that the *MSE* becomes more than two times smaller if the number of studies increases from 10 to 30. If the multilevel analysis involves only 10 studies, the *MSE* is reduced when there are more cases and measurements per case. If the analysis involves 30 studies, the number of cases and the number of measurements hardly influence the *MSE*. This finding is very important in practice: if we want to combine small single-case studies (which is very realistic), the multilevel approach will give us good results if the number of studies is large enough. This finding corresponds to the theoretical expectation and numerical examples of the two-level analysis of group-comparison data (Snijders & Bosker, 1993). But, in some realistic conditions it is infeasible to combine single-case data from 30 studies. We explored conditions where only 10 studies were involved and we found unbiased estimates of the average treatment effects. The *MSE* is small,

the standard error is well estimated, the coverage proportion is close to the nominal value of .95 and the power is large when the studies and the cases are homogeneous ($\sigma^2_{v_2}$ = 0.5 and $\sigma^2_{u_2}$ = 0.5 or $\sigma^2_{u_2}$ = 2). However, the power for the estimate of the treatment effect on the time trend does not reach the threshold of .80 in any condition. The power can be increased by including homogeneous cases, homogeneous studies and including a large number of cases (*J* = 7). Regarding the power when testing the treatment effects, we found that for the effect sizes we tested, which are typical in single-case research, the minimum requirements to obtain a power of .80 for the test of the effect on the time trend were at least 30 studies and homogeneity among the studies (i.e., a small amount of between-study variance). This confirms the theoretical expectation and numerical examples of Snijders and Bosker (1993) who state that the top-level units are of fundamental importance for optimizing power. Fewer studies or more heterogeneity among studies could be tolerated if the interest was just on the test of the immediate effect, or if the true effect size was larger than what we considered.

As with any multilevel model, implausibly large variance components estimates can be obtained. From the estimates of the variance components, we deduce that there are biases when estimating the between-case variance of the immediate treatment effect and the effect on the time trend. These biased variance estimates are consistent with previous empirical research about the three-level analysis of single-case data (Owens and Ferron, 2012) and previous research from a broader methodological domain, for instance growth curve models (Kwok, West, & Green, 2007; Murphy & Pituch, 2009). The bias is more substantial when estimating the between-study variance of the immediate treatment effect and the effect on the time trend. In these contexts we recommend the inclusion of 30 or more studies, and even then researchers should anticipate some bias, particularly when the between-case variance is large compared to the between-study variance. If large variance components estimates are obtained, we advise researchers to respecify the random part of the multilevel model.

In this study, we used a multiple-baseline across participants design, regarded as strong single-case designs because the staggering of the timing of the intervention makes it less plausible to attribute changes in the data to other extraneous events rather than the treatment thereby strengthening the design's internal validity (Shadish et al., 2002). Still, it is recommended to identify the quality of the initial studies before combining them in a three-level analysis, because this quality will immediately affect the quality of the meta-analytic results. To score single-case studies on quality, different instruments can be used, for instance the Single-case Experimental Design (*SCED*) Scale developed by Tate et al. (2008). The

SCED Scale is characterized by high levels of inter-rater reliability. Dependent on the score on quality, researchers can decide to include or not include the study in the three-level analysis or to give a weight to a study's results before including them.

2.4.2 Recommendations for single-subject analysts

The study shows that the average treatment effects are generally well estimated if the between study-variability is small and if a minimum of 30 studies are involved. The number of measurements and cases is of less importance. Researchers can try to measure outcome variables in a consistent way across subjects and even across studies measuring the same constructs. Besides the importance of systematically varying characteristics of studies in order to investigate moderator effects, it also might be advantageous to replicate previous studies, resulting in homogeneous study results. Of course the methodology and instruments have to remain appropriate for the subject in a certain context.

2.4.3 Limitations and suggestions for future research

Although the three-level analysis of single-case experimental studies is promising, some limitations remain. A first limitation is that we can only use a three-level analysis of single-case data, when the raw data are available and on the same scale. Otherwise we depend on standardized effect size estimates and associated meta-analytic procedures. Usually single-case data are presented graphically in the primary studies, which allow us to retrieve the raw data.

The Monte Carlo method used in this study provided control of specific factors (the effect on intercept, the effect on trend, the number of studies, the number of cases, the number of measurements, the between-study variance and the between-case variance) to investigate the appropriateness of inferences made from a three-level single-case model in specific situations. Despite the large number of conditions and the choice of realistic values for the parameters, we still have to be careful with the generalization to conditions that were not simulated.

When estimating the variance components, in some conditions convergence was not reached and the final variance estimates were very large. Therefore we deleted the variance estimates larger than 100 (this was the case in 7.10^{-5} % of the datasets). The variance estimates are still positively skewed after deletion of the extreme values (skewness = 2.38) due to truncation of negative estimates to zero.

In this study, we used a multiple-baseline across participants design in which we staggered the timing of the treatment. Staggering in multiple-baseline designs is often done to enhance the internal validity: if for the participant the scores change at the start of the treatment, it is unlikely that this is due to an external event (Ferron & Onghena, 1996; Ferron & Sentovich, 2002). A limitation of our simulation study is that in simulating the data, we did not account for potential confounding events that could have had a simultaneous effect on all participants.

Despite the strengths of the simulation study, there are several aspects that need further exploration. First, in the three-level analyses, it is assumed that residuals are independent across measurement occasions. However, this assumption of independence may be violated because of autocorrelation (Beretvas & Chung, 2008; Kratochwill et al., 1974). In single-case data, random context variables that influence the score at a certain moment can also influence scores on one or more succeeding occasions which lead to similarity among errors that are close to each other in time (Kromrey & Foster-Johnson, 1996). On the one hand, the issue of autocorrelation cannot be ignored (Ferron et al., 2009; Huitema & McKean, 1994; McKnight, McKean, & Huitema, 2000). Previous research indicates that not modeling existing autocorrelation in a two-level analysis of single-cases results in biased parameter estimates (Ferron et al., 2009). However on the other hand, Shadish and Sullivan (2011) indicated that the size of autocorrelation in SSED studies varies tremendously, with an average of about zero. In this study, we assumed that the repeated measures within a case are independent because we only wanted to evaluate the basic three-level model. In the three-level model we proposed, we modeled the level-one errors as $\sigma^2 I$, but there are many other covariance structures possible, of which the first-order autoregressive type is often used to model autocorrelation (Ferron et al., 2009; McKnight et al., 2000). The current study should be extended, for example, by generating a first-order autoregressive structure.

Secondly, we simulated data assuming (multivariate) normal distributions of the residuals at each level. Further research is needed to investigate the performance of the approach if the normality assumption that is made in the multilevel approach is violated, for instance if the dependent variable is composed of count data.

Third, we used a multiple-baseline across participants design, but there are other designs like alternating treatment designs and reversal designs that need further exploration. Another extension that would be worthwhile to investigate is how a model that ignores a linear or nonlinear time trend would perform if there is a linear or nonlinear trend in the data?

A challenging question is what to do when the obtained data are measured on different scales. Van den Noortgate and Onghena (2003b, 2008) propose as a first possible solution to standardize the data per case by dividing the data by the estimated root mean squared error that is found when using the Center et al. (1985-1986) regression model on the data for that case. Moeyaert et al. (2013b) conducted recently a three-level analysis on standardized single-case data and found that this way of standardizing is appropriate for the estimation of the treatment effects, especially when many studies (30 or more) and a lot of measurements occasions within subjects (20 or more) are included and when the studies are rather homogeneous (with a small between-study variance). The estimates of the variance components are less accurate. A second possible solution is extending the three-level model by including characteristics of the dependent variable as predictors.

This research shows that the three-level synthesis of single-cases works relatively well under a variety of realistic conditions, and especially when the number of studies is large and the studies are homogeneous, but further research is needed to give further answers on the unresolved questions listed above.

Chapter 3|

Three-Level Analysis of Standardized Single-Case Data[2]

Abstract

Previous research indicates that three-level modeling is a valid statistical method to make inferences from unstandardized data from a set of single-subject experimental studies, especially when a homogeneous set of at least 30 studies are included (Moeyaert et al., 2013a). When single-subject data from multiple studies are combined, however, it often occurs that the dependent variable is measured on a different scale, requiring standardization of the data before combining them over studies. One approach is to divide the dependent variable by the residual standard deviation. In this study we use Monte Carlo methods to evaluate this approach. We examine how well the fixed effects (i.e., immediate treatment effect and treatment effect on the time trend) and the variance components (i.e., the between and within-subject variance) are estimated under a number of realistic conditions. The three-level synthesis of standardized single-subject data is found appropriate for the estimation of the treatment effects, especially when many studies (30 or more) and many measurement occasions within subjects (20 or more) are included and when the studies are rather homogeneous (with small between-study variance). The estimates of the variance components are less accurate.

Keywords: single-subject experimental data, three-level analysis, Monte Carlo simulation study, standardized single-subject data

[2] This chapter has been published as Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S.N.,& Van den Noortgate, W. (2013b). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*, 719-748. doi: 10.1080/00273171.2013.816621

3.1 Introduction

In a single-subject experimental design (SSED), the outcome variable of one subject is measured repeatedly within and across different conditions or phases (e.g. baseline phase or A-phase, treatment phase or B-phase). Although the use of SSEDs has grown, systematic reviews and meta-analyses of treatment effects often include only studies using group-comparison studies to estimate changes between different conditions under investigation (Van den Noortgate & Onghena, 2008). The exclusion of SSEDs from these reviews is a matter of concern because information about the variation between subjects in the magnitude of treatment effects tends to be lost in group-comparison designs, which provide averages and effect sizes only for the entire group.

A limitation of single-subject designs is that the corresponding results are subject-specific and therefore not generalizable to other subjects. In order to address this problem, researchers can replicate single-subject experiments within studies (e.g. multiple-baseline designs). Among single-subject designs, these multiple-baseline designs are preferred (Shadish & Sullivan, 2011), because the staggering of the treatment across subjects makes it possible to disentangle real treatment effects from extraneous factors like maturation or history. As a result, these designs are increasingly popular, as shown in Figure 3.1.

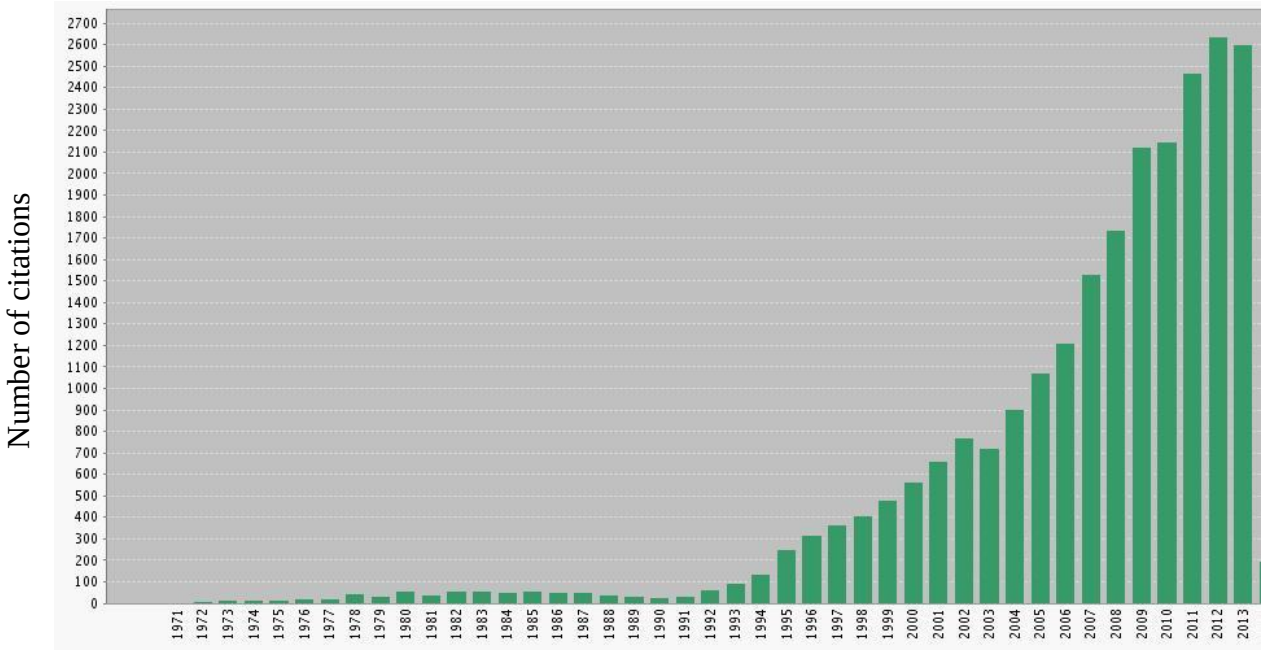

Figure 3.1. Graphical display showing the increase in the number of citation for the keyword "multiple-baseline" between 1971 and 2013 using the Social Science Citation Index within the Web of Sciences.

Another way to address the problem of generalizability is the replication of SSED across studies. Combining data of replicated SSEDs can for instance be accomplished by using a meta-analysis of effect sizes (Busk & Serlin, 1992; Shadish et al., 2012; Rindskopf et al., 2012; Maggin et al., 2011). A problem is that there is no consensus in the literature about the effect size metric to be used. A number of nonparametric effect size metrics have been proposed to analyze single-case designs (e.g., percentage of non-overlapping data, percentage of all non-overlapping data, or percent exceeding the median). Although these nonparametric effect size measures for single-case research can be used without making distributional assumptions, they entail at least three weaknesses. First, such measures may be influenced by outliers in the baseline phase (Allison & Gorman, 1993; Salzberg, Strain, & Baer, 1987). A second drawback is their insensitivity to data trends and variability in the data (White, 1987; Wolery et al., 2010), and third, the sampling distributions of these metrics are unknown, which limits the validity of statistical tests such as moderator analyses that are often conducted in meta-analytic work (Beretvas & Chung, 2008). Other effect size measures are based on regression models (Beretvas & Chung, 2008; Van den Noortgate & Onghena, 2003b). Treatment effects in single-case studies are further sometimes tested using nonparametric randomization tests. Randomization tests can also be used to test the existence of a treatment effect in a set of single-case studies (Edgington & Onghena, 2007). Up to now, the most common way to analyze single-case data is by using visual analyses (Barlow & Hersen, 1984; Kazdin, 1982; Kennedy, 2005; Kratochwill, 1978; Kratochwill & Levin, 1992; Tawney & Gast, 1984). Although visual analysis might do justice to the richness of single-case data, this method tends to result in too many Type I errors (Fisch, 2001; Normand & Bailey, 2006) and Type II errors (Jones, Weinrott, & Vaught, 1978).

In this article we focus on a parametric statistical method to summarize single-case results over cases and over studies, namely the three-level modeling of the raw data. Previous studies indicate that the three-level modeling is a valid statistical method to combine data (Ferron et al., 2010; Moeyaert et al., 2013a; Owens & Ferron, 2012; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003b, 2008).

A complexity when single-subject data from multiple studies are combined is that the dependent variable often is not measured on a common scale, requiring a standardization of the data before combining them over studies. Given the importance of standardization, we discuss and evaluate in this study one specific standardization method, used before combining the data by means of a three-level model. In the following paragraphs, we first present the three-level model to aggregate single-subject data. Then, we describe the method to

standardize single-subject data. Next, we present the setup and results of a Monte Carlo simulation study evaluating the analysis of the three-level modeling of standardized single-subject data.

3.1.1 Three-level modeling

Hierarchical structures occur naturally: for instance, patients are clustered, nested, or grouped within clinics within health authorities, voters are nested within polling districts within constituencies, and citizens are grouped within cities within countries. Kreft and De Leeuw (1998) expressed this as follow: "Once you know that hierarchies exist you see them everywhere" (p. 1). Also data in social and behavioral sciences are usually characterized by a hierarchical structure and therefore require statistical analysis methods that account for this structure. Schooling systems present an obvious example of a hierarchical structure: students are grouped within classes which themselves are grouped within schools. We refer to a hierarchy as consisting of units grouped at different levels. In this example, students are the level-1 units, classes the level-2 units, and the schools the level-3 units in a three-level structure. A different example of hierarchically structured data occurs when the same case or subject is measured repeatedly within and across different conditions or phases (e.g., a baseline phase and a treatment phase), such as in SSEDs. If we have a set of studies in which one or a few subjects are investigated, we can see a three-level structure: measurement occasions at the first level are grouped within cases or subjects at the second level, which in turn are grouped in studies at the third level (see Figure 3.2).

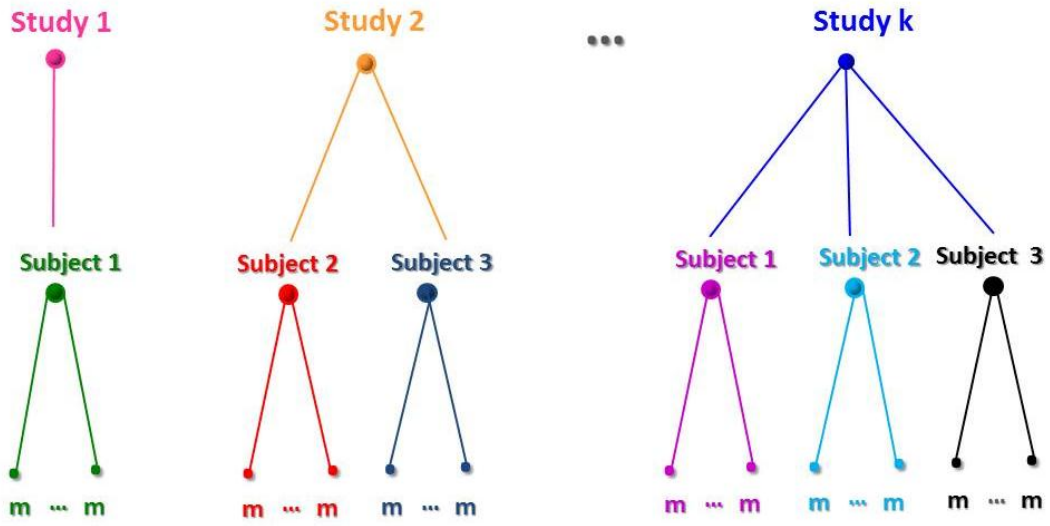

Figure 3.2. The three-level hierarchical structure for the synthesis of single-subject experimental data.

A three-level model can be used to analyze such a data structure. An advantage of the use of a three-level model is that it allows one to estimate within-subject, between-subject, and between-study variance. Moreover, ignoring the study level would imply that we do not take into account that subjects from the same study are more alike than subjects from different studies. Van den Noortgate, Opdenakker, and Onghena (2005) showed that ignoring a top (or intermediate) level has significant effects on the results of a multilevel analysis using hierarchical linear models.

At the first level of this three-level model, a regression equation describes the within-subject variability (Equation 3.1). Y_{ijk} describes the score on the dependent variable on measurement occasion i (i = 1, 2, …, I), for subject j (j = 1, 2, …, J) in study k (k = 1, 2, …, K) as a linear function of two predictors and their interaction, more specifically a time indicator (T_{ijk}), for instance the session number, and a dummy coded variable (D_{ijk}) indicating whether the measurement occasion i from the j^{th} subject in study k belongs to the baseline phase (D_{ijk} = 0) or the treatment phase (D_{ijk} = 1).

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}T_{ijk}D_{ijk} + e_{ijk} \text{ with } e_{ijk} \sim N(0, \sigma_e^2) \tag{3.1}$$

If the time indicator is coded such that it equals zero at the start of the treatment phase, β_{0jk} indicates the expected baseline level at the start of the treatment phase (when T_{ijk} = 0), β_{1jk} is the linear time trend in the baseline scores, the coefficient β_{2jk} is then the immediate effect of the treatment on the outcome, and β_{3jk} refers to the effect of the intervention on the trend. The regression coefficients have indexes j and k, meaning that they are subject- and study-specific.

At the second level, the variation across subjects is modeled in the following four equations:

$$\begin{cases} \beta_{0jk} = \theta_{00k} + u_{0jk} \\ \beta_{1jk} = \theta_{10k} + u_{1jk} \\ \beta_{2jk} = \theta_{20k} + u_{2jk} \\ \beta_{3jk} = \theta_{30k} + u_{3jk} \end{cases} \text{ with } \begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{2jk} \\ u_{3jk} \end{bmatrix} \sim N(0, \Sigma_u) \tag{3.2}$$

These equations indicate that the β coefficients from Equation 3.1 equal an average study-specific performance, the θ coefficients, plus a random variation from these means.

At the third level, the variation of the study-specific regression coefficients from the second level equations is described:

$$\begin{cases} \theta_{00k} = \gamma_{000} + v_{00k} \\ \theta_{10k} = \gamma_{100} + v_{10k} \\ \theta_{20k} = \gamma_{200} + v_{20k} \\ \theta_{30k} = \gamma_{300} + v_{30k} \end{cases} \text{with} \begin{bmatrix} v_{00k} \\ v_{10k} \\ v_{20k} \\ v_{30k} \end{bmatrix} \sim N(0, \Sigma_v) \tag{3.3}$$

Researchers typically are interested in the regression coefficients at the third level (the γ's in Equation 3.3, called fixed effects), especially in γ_{200} and γ_{300} because they represent the average treatment effects (i.e. the immediate treatment effect and the treatment effect on the time trend), as well as in the variance components at each level.

The functioning of the three-level synthesis of unstandardized SSED data was already evaluated by the simulation study of Owens and Ferron (2012) focusing on the most basic multiple-baseline designs, in which there were no trends in either phase. Recently, Moeyaert et al. (2013a) simulated three-level data including time trends. The simulation studies show that the three-level approach results in unbiased estimates of both kinds of treatment effects.

3.1.2 Standardized single-subject experimental data

Dependent variables in a set of SSED-studies are not always measured the same way and on the same scale. For instance, challenging behavior in class in one study is measured on a scale from one to ten, whereas another researcher indicates the challenging behavior on a scale from one to five. Therefore standardization is needed to allow immediate comparison and fair interpretations of scores on challenging behaviour across different studies. In order to standardize raw single-case data, Z-scores could also be used by subtracting the mean from each outcome score, and dividing this difference by the standard deviation. But, in this case we lose important information about the baseline level of the case. When combining the results of multiple cases, we therefore will also not obtain information about the mean baseline level and differences in baseline level between cases. A second option is to divide the outcome score by the standard deviation of the dependent variable (i.e. Z-scores), but this latter depends not only on the scale of the dependent variable, but also on the value of the treatment effects. We do not recommend this method, because the treatment effects (i.e., immediate treatment effect and treatment effect on level) can differ from case to case, and therefore dividing our scores by the standard deviation of the outcome scores could even make effects measured on the same scale not comparable anymore.

Van den Noortgate and Onghena (2008) proposed a method to standardize individual level raw data that addresses all these limitations. They proposed to perform an ordinary least squares (OLS) regression for each subject from one study separately (for instance using Equation 3.1) in order to estimate the residual within-subject standard deviation ($\hat{\sigma}_{ejk}$). Thereafter the individual scores ($Y_{ijk}{}'s$) are divided by the estimated residual within-subject standard deviation ($\hat{\sigma}_{ejk}$):

$$Y'_{ijk} = \frac{Y_{ijk}}{\hat{\sigma}_{ejk}} \tag{3.4}$$

The rationale behind this method is that in many situations, the scale of the dependent variable in one study likely differs from the scale used in another study, for instance in a scenario where behavior in one study is rated using a 5-point scale rather than on a 7-point scale as in another study. This means that results will be on different scales too. The residual within-subject standard deviation reflects these kinds of differences in how the dependent variable is measured, and thus dividing the original raw scores in a study by this variability provides a method of standardizing the scores. At the same time, it is not impacted by the size of the treatment effect, and thus is not expected to bias the treatment effect estimates. We attached an empirical illustration of this standardizing method in Addendum A1 (step 1 through step 4). Thereafter, the standardized data can be combined over subjects and over studies using Equations 3.1 through 3.3. The aim of this simulation study is to evaluate whether the three-level model approach is still appropriate if single-subject data are standardized using Equation 3.4.

3.2 Simulation Study

Simulation studies allow us to control the population values and to assess the validity, accuracy and power of statistical procedures in realistic but generic situations. First, we simulated raw single-subject data using Equation 3.1 to 3.3. Next, we estimated for each subject the unstandardized regression coefficients of Equation 3.1, using OLS estimation. In a following step, we used Equation 3.4 to obtain standardized outcome scores. More specifically, for each subject we divided the outcome scores (i.e., Y_{ijk}'s), by the residual within-phase standard deviation (Equation 3.4). Finally, the standardized data (i.e., Y'_{ijk}'s) were analyzed using the three-level approach and results were compared to the parameter values used to generate data. To estimate the three-level model parameters, the restricted

maximum likelihood procedure in SAS 9.3 PROC MIXED was used (Littell et al., 2006). To estimate the degrees of freedom, we used the Satterthwaite approach because this method seems to provide accurate confidence intervals for the estimates of the average treatment effect for two-level analysis of single-subject data (Ferron et al., 2009).

To evaluate the three-level approach for standardized data, we calculated the bias and mean squared error of the effect parameter estimates, the corresponding standard error estimates, and the coverage proportion of the 95% confidence intervals for the fixed effects. Moreover, we focused on the power for testing treatment effects. Next, we calculated the relative bias of the variance components.

Furthermore, the following seven parameters were varied: The number of simulated measurements within subjects, the number of subjects per study, the number of studies, the immediate treatment effect, the treatment effect on the time trend, the between-subjects variance, and the between-study variance. A description of the levels chosen for each of these factors is provided below, along with justification for the levels selected for study.

The total number of simulated measurements within a subject was equal to 10, 20, or 40 (I = 10, 20 or 40). We chose to keep I constant for all subjects within and across studies. The values for the measurements were selected based on a survey of multiple-baseline studies (Ferron et al., 2010), which found average measurement occasions with a median of 24, and based on a meta-analysis of 85 single-subject studies (Swanson & Sachse-Lee, 2000), which found that 25 studies had fewer than 11 treatment measurements, 37 studies had between 11 and 29 treatment measurements, and 23 studies had more than 29 treatment measurements. Shadish and Sullivan (2011) found a median number of measurements of 20, and identified that 90.6% of the subjects had 49 or fewer measurement occasions.

The number of subjects per study equaled 3, 4 or 7 (J = 3, 4 or 7). These values were chosen based on a survey of multiple-baseline studies (Ferron et al., 2010), which included multiple-baseline studies having a median of 4 subjects, based on recommendations of Barlow and Hersen (1984) to include three subjects and Kazdin and Kopel (1975) to include four baselines. Moreover Shadish and Sullivan (2011) reported the characteristics of 809 single-subject studies and found that the number of subjects per study ranged from 1 to 13 with a median of 3. Farmer et al. (2010) reported in their review that 93% of the average number of subject per study was equal to or less than 7.

The number of simulated studies was 10 or 30 (K = 10 or 30). A review of social science single-subject meta-analysis (Farmer et al., 2010) showed that the number of studies included in a meta-analysis ranged from 3 to 117, with 60% of the meta-analysis including

less than 30 studies. We chose to include only lower limits for the number of studies (K = 10 or 30) to test if the model works appropriately in these conditions.

The immediate treatment effect, γ_{200}, had values of 0 (no effect) or 2, and the treatment effect on the time trend, γ_{300}, equaled 0 (no effect) or 0.2. This is based on several re-analyses of meta-analyses (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011). The regression coefficients of the baseline γ_{000} and γ_{100} were not varied (and are set at 0), because the focus of the current study is on treatment effects.

The between-subject variance, Σ_u, had conditions with relatively large and small amount of between-subject variance and sometimes greater than the within-person variance (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011): $\Sigma_u =$ diag($\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}$) = diag(2, 0.2, 2, 0.2), diag(0.5, 0.05, 0.5, 0.05), and diag(8, 0.08, 8, 0.08). Again based on re-analyses of meta-analyses (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011), we have chosen the same sets of values for the four diagonal elements of the between-study variance: $\Sigma_v =$ diag($\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}$) = diag($2, 0.2, 2, 0.2$), $\Sigma_v =$ diag($\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}$) = diag($0.5, 0.05, 0.5, 0.05$), and $\Sigma_v =$ diag($\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}$) = diag($8, 0.08, 8, 0.08$).

We choose to keep the within-subject variability (standard deviation) constant across subjects, meaning that the simulated data are on the same scale for each subject and study. If we had simulated data using different scales, this effect would be compensated by the standardization (e.g., if for a specific subject we had multiplied each score by three, the estimated residual standard deviations would be 3 times larger, so the estimated standardized scores would remain unchanged). By simulating data on the same scale, it is possible to evaluate at the same time the multilevel approach for SSED analysis without standardization and the approach with standardization.

Because we simulated data using a multiple-baseline across participants design, we staggered the introduction of the intervention across subjects within studies. The staggering depended on the total number of measurement occasions and the number of subjects (see Table 3.1).

Table 3.1

Starting Moment of the Intervention as a Function of the Number of Subjects (J) and the Number of Measurements (I)

J		*I* = 10	*I* = 20	*I* = 40
3	Subject 1	4	7	11
	Subject 2	6	11	21
	Subject 3	8	15	31
4	Subject 1	4	7	11
	Subject 2	5	10	18
	Subject 3	7	12	24
	Subject 4	8	15	31
7	Subject 1	4	7	11
	Subject 2	5	9	15
	Subject 3	5	9	15
	Subject 4	6	11	21
	Subject 5	7	13	27
	Subject 6	7	13	27
	Subject 7	8	15	31

Crossing the levels of the seven varying factors, leads to a 3x3x2x2x2x3x3 factorial design including 648 experimental conditions. For each condition, we simulated 2,000 replications resulting in a total of 1,296,000 datasets. In order to analyze the variation between the experimental conditions for both treatment effects and the variance components, we used the procedure PROC GLM in SAS 9.3. The dependent variables were: the deviations of the estimated average treatment effects from the true population values, the squared deviation, the difference between the estimated standard error and the standard deviation of the treatment effects, the coverage proportion of the 95% confidence interval, the power, and the deviations of the variance estimates from their true population values. The experimental conditions that have a statistically significant ($p < .001$) effect on the above listed dependent variables were further explored. This procedure was only used as a preliminary investigation to discover the conditions that can play a significant role in the three-level synthesis of single-subject data.

3.3 Results of the Simulation Study

The estimated average treatment effects (i.e., the immediate treatment effect and the treatment effect on the time trend) are discussed in the first section. In the second section we present the relative bias of the point estimates of the variance components (i.e., the between-subject within study variance and the between-study variance). The full dataset is available from the first author.

3.3.1 Average treatment effect

3.3.1.1 Bias and mean squared error

Previous simulation studies evaluating the three-level analysis of unstandardized single-subject data indicate that the estimated relative bias for both estimated treatment effects was close to zero (Moeyaert et al., 2013a; Owens & Ferron, 2012). We also expected this in current study, which was confirmed. When γ_{200} and γ_{300} equal 0, the bias was respectively 0.00079 and -0.000027 and when γ_{200} = 2 and γ_{300} = 0.2, the relative biases were respectively 0.074 and 0.075. The number of measurements had a substantial influence on the relative bias. Especially if the number of measurements is small (I = 10), the relative bias is substantial if γ_{200} = 2. The relative bias for 10 measurements is .15 whereas the relative bias for 20 measurements equals .05 (see Figure 3.3). The bias is further reduced by adding more measurements. When γ_{200} = 0, the bias is close to zero, independent of the number of measurements. These conclusions are similar for the relative bias of the estimated treatment effect on the slope.

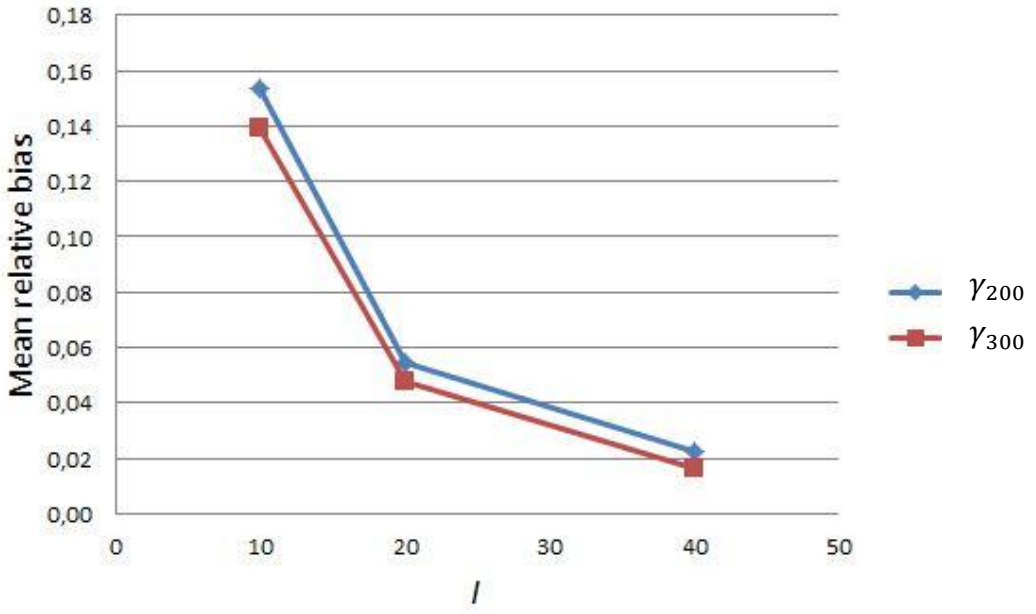

Figure 3.3. Influence of the number of measurements on the relative bias of the estimated immediate treatment effect and estimated treatment effect on time trend; for $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $K = 10$, $J = 4$, $\sigma^2_{u_0} = 2$ and $\sigma^2_{v_0} = 2$ conditions.

Furthermore, we sorted all conditions by their relative bias for the estimate of both treatment effects. When estimating the immediate treatment effect, the relative bias is largest (3.11) in the condition with: $\gamma_{200} = 2$, $\gamma_{300} = 0$, $K = 10$, $J = 3$; $I = 10$, $\sigma^2_{v_2} = 8$ and $\sigma^2_{u_2} = 8$. This is comparable to the condition in which the bias when estimating the treatment effect on the time was largest (4.53): $\gamma_{200} = 0$, $\gamma_{300} = 0.2$, $K = 10$, $J = 3$; $I = 10$, $\sigma^2_{v_2} = 0.2$ and $\sigma^2_{u_2} = 0.05$. The relative bias in these conditions can significantly be reduced by including at least 20 measurement occasions per subject: when going from 10 to 20 measurement per subject, the mean relative bias of the immediate treatment effect and the treatment effect on the time trend are respectively: 0.048 and 0.062.

We estimated the Mean Squared Error (*MSE)* of both estimated treatment effects because it gives information about their bias and variance around the population effect. Table 3.2 provides the *MSE* for the immediate effect of the treatment (γ_{200}) where $\gamma_{200} = 2$ and $\gamma_{300} = 0.2$. Similar patterns are seen for the other combinations of treatment effect values.

As expected, the *MSE* for both treatment effects, γ_{200} and γ_{300} becomes smaller when the number of studies increases from 10 to 30. The *MSE* is further affected by the size of the between-study and the between-subject variance (see Table 3.2). The larger the between-subject and especially the between-study variance, the larger the *MSE*. The *MSE* is influenced to a much smaller degree by the number of subjects. The *MSE* in this study is also only decreasing slightly with an increasing number of measurements. This influence is mostly visible when only 10 studies are involved. If the values in Table 3.2 are compared to the *MSE*

values from the unstandardized data (available in Moeyaert et al., 2013a), it can be seen that the *MSE* for standardized and unstandardized data are comparable if there are at least 20 measurements within a subject, but the *MSE* for standardized data remains slightly higher. The conclusions are very similar for the estimates of γ_{300}.

Table 3.2

Mean Squared Error of the Estimated Immediate Treatment Effect; for γ_{200} = 2 and γ_{300} = 0.2 Conditions

			K = 10			*K* = 30		
I	*J*	$\sigma_{u_2}^2$	$\sigma_{v_2}^2$ = 0.5	$\sigma_{v_2}^2$ = 2	$\sigma_{v_2}^2$ = 8	$\sigma_{v_2}^2$ = 0.5	$\sigma_{v_2}^2$ = 2	$\sigma_{v_2}^2$ = 8
10	3	0.5	0.12	0.24	0.63	0.09	0.12	0.21
		2	0.17	0.25	0.61	0.09	0.12	0.26
		8	0.28	0.35	0.75	0.14	0.18	0.28
	4	0.5	0.13	0.20	0.52	0.10	0.11	0.22
		2	0.15	0.22	0.64	0.09	0.11	0.24
		8	0.24	0.33	0.64	0.13	0.14	0.28
	7	0.5	0.11	0.18	0.50	0.09	0.11	0.23
		2	0.12	0.21	0.55	0.09	0.11	0.24
		8	0.19	0.28	0.64	0.10	0.12	0.23
20	3	0.5	0.05	0.12	0.41	0.02	0.05	0.15
		2	0.08	0.14	0.49	0.03	0.06	0.16
		8	0.19	0.26	0.51	0.06	0.09	0.20
	4	0.5	0.05	0.13	0.44	0.02	0.05	0.16
		2	0.07	0.15	0.46	0.03	0.05	0.16
		8	0.16	0.21	0.53	0.05	0.07	0.17
	7	0.5	0.04	0.12	0.42	0.02	0.04	0.15
		2	0.05	0.13	0.44	0.02	0.04	0.15
		8	0.10	0.16	0.48	0.03	0.06	0.17
40	3	0.5	0.04	0.11	0.37	0.01	0.04	0.13
		2	0.07	0.15	0.43	0.02	0.05	0.14
		8	0.15	0.23	0.54	0.05	0.08	0.17
	4	0.5	0.04	0.11	0.35	0.01	0.03	0.12
		2	0.05	0.12	0.45	0.02	0.04	0.14
		8	0.14	0.21	0.48	0.04	0.06	0.17
	7	0.5	0.03	0.11	0.35	0.01	0.03	0.12
		2	0.04	0.11	0.41	0.02	0.04	0.13
		8	0.09	0.14	0.45	0.03	0.05	0.15

3.3.1.2 Estimates of standard errors of the average treatment effects

In this section, we evaluate the estimate of the standard errors that can be used to construct confidence intervals. Standard errors are per definition equal to the standard deviation of the sampling distribution. Therefore, we can use the standard deviation of the treatment effect estimates in a specific condition as an approximation of the true standard errors, and use this standard deviation as a criterion to evaluate the estimated standard errors. The larger the number of estimates of the effects, the better the standard deviation of the estimates is expected to correspond to the true standard error.

The relative difference between the median standard error estimates and the standard deviation of the estimates of the fixed effects was obtained by dividing the difference by the standard deviation of the estimates. The relative difference tends to be negative (see Table 3.3), which indicates that the standard errors are underestimated, however none of the bias values found in the scenarios examined here exceeded Hoogland and Boomsma's (1998) criterion value of 10% (1,000 out of 10,000). Because some of the values approached the cutoff, we still examined which factors affected the relative standard error bias. The measurement occasions, $F(2,645) = 15.74$, $p < .001$, the number of studies, $F(1,646) = 302.85$, $p < .001$, and the between-subject variance , $F(2,645) = 15.92$, $p < .001$, were found to have a significant impact on the gap between the standard error estimates and the standard deviation of the estimates (see Table 3.3). To reduce the gap, one can increase the number of measurements and increase the number of studies. For example, increasing the number of studies to 30 results in a better estimate of the standard error of the immediate treatment effect. Similar results were found for the synthesis of unstandardized data, except that the number of measurements does not matter for unstandardized data. When we estimate the treatment effect on the time trend, γ_{300}, the same conclusions can be made.

Table 3.3

*Relative Difference (*10,000) between the Median of the Standard Error Estimates and the Standard Deviation of the Estimated Immediate Treatment; for γ_{200} = 2 and γ_{300} = 0.2 Conditions*

			K = 10			*K* = 30		
I	*J*	$\sigma^2_{u_2}$	$\sigma^2_{v_2} = 0.5$	$\sigma^2_{v_2} = 2$	$\sigma^2_{v_2} = 8$	$\sigma^2_{v_2} = 0.5$	$\sigma^2_{v_2} = 2$	$\sigma^2_{v_2} = 8$
10	3	0.5	-777	-668	-641	-580	-405	32
		2	-497	-787	-650	-353	-166	-291
		8	-69	-531	-446	104	-318	-145
	4	0.5	-826	-619	-404	-252	-241	-377
		2	-565	-198	-558	-256	-129	-99
		8	-141	-213	-398	-195	24	-267
	7	0.5	-498	-491	-365	-74	-198	-215
		2	-407	-698	-274	-91	-15	-66
		8	-409	-615	-495	-218	-195	41
20	3	0.5	-433	-601	-281	-233	-261	29
		2	-159	-288	-564	62	-83	-221
		8	-82	-308	-150	-83	-380	-103
	4	0.5	-285	-502	-424	57	-181	-210
		2	-581	-231	-581	-186	459	-311
		8	-239	-42	-151	-54	105	92
	7	0.5	-200	-418	-370	-140	-508	-145
		2	-264	-461	-479	-179	101	-24
		8	-171	-229	-610	17	-256	-324
40	3	0.5	-493	-526	-355	-216	47	-261
		2	-491	-479	-430	-148	-105	-116
		8	31	-520	-538	131	-201	-228
	4	0.5	-153	-439	-277	189	132	128
		2	-384	-131	-346	-16	-197	-43
		8	-19	-272	-639	9	-123	-87
	7	0.5	-187	-552	29	-173	174	7
		2	-482	-355	-484	40	-259	-104
		8	-445	-234	-465	11	16	-24

3.3.1.3 Coverage proportion

We calculated 95% interval estimates of the fixed effects, based on the point estimates, the standard error estimates, and the Satterthwaite estimated degrees of freedom. We evaluated these interval estimates by estimating their coverage proportion. An estimated coverage proportion of 95% means that in 95% of the confidence intervals calculated for a specific condition, the population value is included in the confidence interval. Because we simulated 2,000 datasets for each condition, the coverage proportions are expected to be close to the nominal value of .95: the standard error is only 0.005 (i.e., $\sqrt{(0.95 * 0.05)/2{,}000}$). The lower and upper limits of the confidence intervals around the population effect (γ_{200} and γ_{300}) are constructed by multiplying the estimated standard error with the critical value of the *t*-distribution, and subtracting from and adding this product to the estimated value of the effect.

In this study, the maximum value for the coverage proportion of the estimates of the average immediate treatment effect (γ_{200}) did not exceed .97, but the coverage proportion went down to problematically small values (e.g., .69). For the estimation of γ_{300}, the smallest value for the coverage proportion is .92. γ_{200} and γ_{300} both have a median coverage proportion of .95 with a standard deviation over conditions of respectively .029 and .0065. Therefore the deviations of the coverage proportion from the nominal value of .95 for the estimation of γ_{200} can not simply be explained by chance in contrast to the analysis on unstandardized data (Moeyaert et al., 2013a).

Table 3.4 presents the coverage proportion for the estimate of γ_{200}, for γ_{200} = 2 and $\gamma_{300} = 0.2$. Similar results are obtained for the combination of the other values for the treatment effects. The coverage proportions are especially small when there are 30 studies with only 10 measurement occasions within a subject. An explanation is that if only 10 measurements per subject are made, the effect is underestimated. This will especially be visible if we have a large number of studies, because in this case the standard error and therefore the confidence interval is smallest.

Table 3.4

Coverage Proportion for the Estimated Immediate Treatment Effect; for γ_{200} = 2 and γ_{300} = 0.2 Conditions

			K = 10			*K* = 30		
I	*J*	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8
10	3	0.5	.92	**.93**	**.95**	.79	.88	**.94**
		2	**.93**	**.93**	**.94**	.84	.90	**.94**
		8	**.95**	**.95**	**.94**	**.93**	.92	**.94**
	4	0.5	.90	**.93**	**.94**	.75	.87	.91
		2	.92	**.95**	**.94**	.81	.89	.93
		8	**.95**	**.94**	**.94**	.89	.92	**.94**
	7	0.5	.89	**.93**	**.94**	.69	.86	.92
		2	.90	**.93**	**.95**	.75	.87	**.93**
		8	**.93**	**.93**	**.95**	.85	.90	**.93**
20	3	0.5	**.94**	**.95**	**.96**	.92	**.94**	**.95**
		2	**.95**	**.95**	**.95**	**.94**	**.94**	**.94**
		8	**.96**	**.96**	**.96**	**.95**	**.94**	**.94**
	4	0.5	**.95**	**.94**	**.95**	.92	**.94**	**.94**
		2	**.94**	**.95**	**.95**	.92	**.95**	**.95**
		8	**.96**	**.95**	**.95**	**.95**	**.94**	**.96**
	7	0.5	**.95**	**.94**	**.95**	.91	.92	**.94**
		2	**.94**	**.95**	**.95**	.92	**.94**	**.95**
		8	**.96**	**.95**	**.94**	**.94**	**.94**	**.94**
40	3	0.5	**.94**	**.95**	**.95**	**.94**	**.95**	**.94**
		2	**.95**	**.95**	**.95**	**.95**	**.95**	**.95**
		8	**.96**	**.95**	**.95**	**.96**	**.95**	**.95**
	4	0.5	**.95**	**.96**	**.95**	**.95**	**.95**	**.96**
		2	**.95**	**.95**	**.95**	**.94**	**.95**	**.95**
		8	**.96**	**.95**	**.95**	**.96**	**.95**	**.95**
	7	0.5	**.95**	**.95**	**.95**	**.94**	**.95**	**.95**
		2	**.95**	**.95**	**.95**	**.95**	**.94**	**.95**
		8	**.95**	**.95**	**.94**	**.95**	**.95**	**.96**

Note. Coverage proportions between .93 and .96 are in boldface.

3.3.1.4 Power

The probability of rejecting the null hypothesis when in fact a certain alternative parameter value is true, is called the power of a significance test (Cohen, 1988). Power calculations can give researchers important information about the minimum required number of units to include in the research design.

We want the power as high as possible when the null hypothesis is false (Cohen, 1988), i.e., when γ_{200} is 2 rather than 0. For the multilevel analysis of standardized single-subject data, the conditions in which the power level is reasonably large (at least .80; Cohen, 1988) are marked in bold (see Table 3.5).

A striking finding is that the power for the estimated immediate treatment effect is always too small when only 10 heterogeneous (e.g. $\sigma_{v_2}^2$ = 8) studies are involved in the three-level analysis. Moreover, when 30 studies are included, the power is above .85 for all conditions. This conclusion is similar for unstandardized data (Moeyaert et al., 2013a).

The results for the power of the estimated treatment effect on the trend, γ_{300}, are also comparable for unstandardized and standardized data (see Table 3.5). The power is too small in all conditions when only 10 studies are involved. When including 30 studies, the power becomes larger, but stays too small if there are a small number of measurements (I = 10) and a small number of subjects (J = 3). When the number of subjects is 4 or 7, the power remains too small if the studies are heterogeneous ($\sigma_{v_3}^2$= 0.2). When the number of measurements becomes larger (I = 20 or I = 40) the power level of .80 is reached in all conditions for rather homogeneous studies ($\sigma_{v_3}^2$= 0.05 or $\sigma_{v_3}^2$= 0.08).

Table 3.5

Power for the Estimated Immediate Treatment Effect and the Estimated Treatment Effect on the Slope; for γ_{200}=2 and γ_{300} = 0.2 Conditions

			$\hat{\gamma}_{200}$							$\hat{\gamma}_{300}$					
			K = 10			*K* = 30				*K* = 10			*K* = 30		
			$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8		$\sigma^2_{v_3}$ = 0.05	$\sigma^2_{v_3}$ = 0.08	$\sigma^2_{v_3}$ = 0.2	$\sigma^2_{v_3}$ = 0.05	$\sigma^2_{v_3}$ = 0.08	$\sigma^2_{v_3}$ = 0.2
I	J	$\sigma^2_{u_2}$							$\sigma^2_{u_3}$						
10	3	0.5	**1.00**	**.92**	.46	**1.00**	**1.00**	**.94**	0.05	.31	.27	.19	.77	.68	.47
		2	**.98**	**.86**	.47	**1.00**	**1.00**	**.92**	0.08	.28	.23	.17	.72	.66	.45
		8	.80	.65	.37	**1.00**	**.99**	**.85**	0.2	.22	.19	.16	.61	.54	.39
	4	0.5	**1.00**	**.94**	.47	**1.00**	**1.00**	**.95**	0.05	.37	.28	.18	**.84**	.75	.50
		2	**.99**	**.90**	.49	**1.00**	**1.00**	**.94**	0.08	.34	.29	.18	**.81**	.71	.48
		8	**.88**	.74	.40	**1.00**	**1.00**	**.89**	0.2	.26	.24	.17	.70	.62	.45
	7	0.5	**1.00**	**.96**	.47	**1.00**	**1.00**	**.96**	0.05	.45	.34	.21	**.93**	**.84**	.55
		2	**1.00**	**.93**	.49	**1.00**	**1.00**	**.95**	0.08	.42	.35	.20	**.92**	**.82**	.55
		8	**.98**	**.85**	.45	**1.00**	**1.00**	**.94**	0.2	.36	.31	.19	**.86**	.77	.52
20	3	0.5	**1.00**	**.93**	.51	**1.00**	**1.00**	**.96**	0.05	.52	.39	.20	**.96**	**.88**	.58
		2	**1.00**	**.91**	.46	**1.00**	**1.00**	**.93**	0.08	.46	.37	.21	**.95**	**.87**	.58
		8	**0.84**	.70	.39	**1.00**	**1.00**	**.89**	0.2	.35	.29	.19	**.81**	.73	.52
	4	0.5	**1.00**	**.96**	.51	**1.00**	**1.00**	**.96**	0.05	.57	.40	.22	**.98**	**.91**	.60
		2	**1.00**	**.94**	.49	**1.00**	**1.00**	**.95**	0.08	.50	.39	.21	**.96**	**.89**	.60
		8	**.92**	.79	.41	**1.00**	**1.00**	**.91**	0.2	.37	.30	.20	**.89**	**.81**	.55
	7	0.5	**1.00**	**.97**	.50	**1.00**	**1.00**	**.96**	0.05	.62	.44	.23	**.99**	**.93**	.63
		2	**1.00**	**.94**	.50	**1.00**	**1.00**	**.96**	0.08	.59	.45	.24	**.99**	**.92**	.63
		8	**.99**	**.86**	.46	**1.00**	**1.00**	**.94**	0.2	.49	.38	.23	**.95**	**.88**	.59
40	3	0.5	**1.00**	**.95**	.50	**1.00**	**1.00**	**.96**	0.05	.58	.42	.24	**.98**	**.93**	.62
		2	**1.00**	**.92**	.46	**1.00**	**1.00**	**.95**	0.08	.52	.41	.24	**.96**	**.88**	.61
		8	**.86**	.73	.39	**1.00**	**1.00**	**.90**	0.2	.36	.30	.20	**.86**	.79	.55
	4	0.5	**1.00**	**.96**	.53	**1.00**	**1.00**	**.96**	0.05	.59	.47	.21	**.99**	**.93**	.61
		2	**1.00**	**.93**	.49	**1.00**	**1.00**	**.96**	0.08	.55	.43	.22	**.98**	**.91**	.62
		8	**.93**	.79	.41	**1.00**	**1.00**	**.91**	0.2	.41	.35	.20	**.92**	**.84**	.58
	7	0.5	**1.00**	**.97**	.50	**1.00**	**1.00**	**.96**	0.05	.66	.49	.24	**.99**	**.95**	.64
		2	**1.00**	**.96**	.51	**1.00**	**1.00**	**.95**	0.08	.60	.46	.26	**.99**	**.94**	.65
		8	**.99**	**.88**	.47	**1.00**	**1.00**	**.94**	0.2	.50	.41	.21	**.97**	**.89**	.60

Note. Values ≥ .80 are in boldface.

3.3.2 Variance components

In addition to the estimation of the fixed effects, we estimated the variance components (i.e., the between-study and the between-subject variances) for both treatment effects (i.e., the immediate treatment effect and the treatment effect on the time trend). We examined the relative bias which is the absolute bias divided by the population parameter. Because the population values for the variance components differ from zero, it is possible to calculate the relative bias in all conditions.

The distribution of the estimated variance components is positively skewed due to transformation of negative estimates to zero. Therefore, we calculated the median relative deviation of the estimates from the population value, rather than the mean relative deviation, to evaluate the relative bias in the estimates.

There is a substantial mean relative bias for the estimate of the between-study variance, 0.24, and the between-subject variance, 0.71, for the immediate treatment effect. Table 3.6 provides the relative bias estimates of the between-case variance and the between-study variance of the immediate treatment effect per condition for y_{200} = 2 and y_{300} = 0.

Similar conclusions can be made when estimating the between-study and the between-subject variance for the effect on the trend, except that the relative biases are smaller. The condition representing the maximum relative bias (0.35) for the estimation of the between-study variance is: K = 30, J = 3, I = 10, $\sigma^2_{v_3}$= 0.05 and $\sigma^2_{u_3}$= 0.02. The condition with the maximum relative bias (2.19) for the estimation of the between-subject variance is K = 30, J = 3, I = 10, $\sigma^2_{v_3}$ = 0.08 and $\sigma^2_{u_3}$ = 0.05.

Table 3.6

Median of Relative Deviation of the Variance Estimates of y_{200}; for y_{200}=2 and y_{300}=0.2 Conditions

			$\hat{\sigma}^2_{v_2}$							$\hat{\sigma}^2_{u_2}$					
			K = 10			*K* = 30				*K =10*			*K = 30*		
			$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}=2$	$\sigma^2_{v_2}=8$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}=2$	$\sigma^2_{v_2}=8$		$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}=2$	$\sigma^2_{v_2}=8$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}=2$	$\sigma^2_{v_2}=8$
I	*J*	$\sigma^2_{u_2}$							$\sigma^2_{u_3}$						
10	3	0,5	0.10	0.20	0.20	0.29	0.31	0.27	0.05	1.92	1.65	3.69	2.26	1.75	4.50
		2	0.09	0.20	0.18	0.30	0.28	0.29	0.08	0.79	0.78	1.36	0.97	1.09	1.58
		8	-0.26	0.08	0.20	0.07	0.25	0.30	0.2	0.41	0.44	0.65	0.57	0.61	0.72
	4	0,5	0.13	0.20	0.20	0.32	0.29	0.27	0.05	2.03	2.45	3.93	2.34	2.88	4.46
		2	0.05	0.21	0.18	0.25	0.30	0.28	0.08	0.85	0.99	1.33	1.02	1.13	1.56
		8	-0.21	0.16	0.16	0.26	0.22	0.27	0.2	0.49	0.54	0.66	0.58	0.63	0.72
	7	0,5	0.22	0.23	0.24	0.31	0.31	0.30	0.05	2.18	2.63	4.24	2.49	2.94	4.85
		2	0.19	0.20	0.22	0.30	0.30	0.31	0.08	0.98	1.06	1.51	1.04	1.15	1.69
		8	-0.06	0.09	0.21	0.26	0.28	0.29	0.2	0.55	0.60	0.71	0.61	0.66	0.77
20	3	0,5	-0.04	-0.02	0.01	0.05	0.06	0.08	0.05	0.50	0.55	0.97	0.58	0.67	1.16
		2	-0.07	-0.04	-0.01	0.07	0.08	0.10	0.08	0.15	0.21	0.32	0.26	0.27	0.41
		8	-0.37	-0.07	-0.01	0.04	0.05	0.07	0.2	0.05	0.10	0.15	0.12	0.16	0.19
	4	0,5	-0.03	0.00	0.01	0.07	0.08	0.07	0.05	0.52	0.64	1.06	0.60	0.73	1.17
		2	-0.09	-0.02	-0.02	0.03	0.09	0.07	0.08	0.21	0.26	0.35	0.26	0.28	0.40
		8	-0.24	-0.05	0.02	-0.04	0.03	0.07	0.2	0.10	0.13	0.16	0.15	0.17	0.20
	7	0,5	0.03	0.01	0.02	0.07	0.07	0.06	0.05	0.57	0.70	1.12	0.61	0.74	1.20
		2	-0.06	0.03	0.01	0.05	0.07	0.08	0.08	0.24	0.28	0.38	0.27	0.30	0.42
		8	-0.15	-0.03	0.00	0.00	0.06	0.06	0.2	0.14	0.15	0.20	0.17	0.18	0.20
40	3	0,5	-0.07	-0.05	-0.04	0.01	0.02	0.01	0.05	0.14	0.23	0.34	0.19	0.24	0.41
		2	-0.14	-0.10	-0.04	0.00	0.01	0.02	0.08	0.05	0.07	0.11	0.08	0.11	0.15
		8	-0.43	-0.17	-0.08	-0.07	0.02	0.00	0.2	-0.02	0.01	0.06	0.04	0.06	0.06
	4	0,5	0.00	-0.07	-0.05	0.03	0.01	0.03	0.05	0.17	0.21	0.36	0.23	0.28	0.44
		2	-0.10	-0.06	-0.04	0.00	0.00	0.03	0.08	0.06	0.09	0.12	0.10	0.11	0.16
		8	-0.37	-0.07	-0.07	-0.08	0.00	0.00	0.2	0.01	0.03	0.05	0.06	0.07	0.08
	7	0,5	-0.5	-0.05	-0.02	0.01	0.01	0.02	0.05	0.19	0.24	0.42	0.22	0.27	0.45
		2	-0.08	-0.03	-0.03	0.00	0.02	0.02	0.08	0.07	0.09	0.13	0.09	0.11	0.16
		8	-0.23	-0.07	-0.05	-0.06	0.01	0.02	0.2	0.04	0.06	0.08	0.06	0.07	0.08

3.4 Empirical Illustration

In this section, we give an empirical illustration of the comparison of the three-level analysis of unstandardized single-subject data and the three-level analysis of standardized single-subject data. Therefore, we selected studies using a multiple-baseline across participants design from the meta-analyses of single-case studies by Heyvaert, Saenen, Maes, and Onghena (2012) in which restraint interventions for challenging behavior among persons with intellectual disabilities was investigated. In total, we retrieved the raw data of 8 studies having 2 to 4 subjects. We combined the studies of Lindberg, Iwata and Kahng (1999); Roscoe, Iwata and Goh (1998); Thompson, Iwata, Conners, and Roscoe (1999); Roane, Piazza, Sgro, Volkert, and Anderson (2001); McCord, Grosser, Iwata, and Powers (2005); Rolider, Williams, Cummings, and Van Houten (1991); Hanley, Iwata, Thompson, and Lindberg (2000); and Zhou, Goff, and Iwata (2000). The SAS codes used for the estimation of the treatment effects over subjects and over studies are described in Addendum A1.

The average immediate treatment effect was -31.24, $t(7) = -3.29$, $p = .01$ and the average treatment effect on the time trend was -0.51, $t(7) = -0.52$, $p = .19$ when we used unstandardized data. When first standardizing the raw data using Equation 3.4 before estimating the effects over subjects and over studies, the immediate treatment effect was -5.14 $t(7) = -3.16$, $p = .01$ and the treatment effect on the time trend was -0.06, $t(7) = -2.92$, $p = .02$. These results indicate that there is a large difference between unstandardized and standardized data with regard to the estimated treatment effect on the time trend. This difference is not surprising because in one study the challenging behavior is counted in intervals that are six times smaller in a study than in another study. For instance, in the study of Roane et al. (2001), the dependent variable is the number of challenging behavior in 10 seconds, whereas in the study of Roscoe et al. (1998), the same challenging behavior is measured as the number of responses per minute. If we standardize the data, outcome scores within phases across studies are closer to each other, resulting in a smaller estimated standard error of the treatment effect on the time trend, $SE(\hat{\gamma}_{300})$. A smaller estimated standard error results in a larger t value (i.e., $t = \frac{\hat{\gamma}_{300}}{SE(\hat{\gamma}_{300})}$) and therefore a significant treatment effect on the time trend across studies is obtained. This means that a significant treatment effect on the time trend would not have been identified if unstandardized data were used. There is also a large difference in the estimated between-subject and between-study variance for both estimated treatment effects. The between-subject variance for the estimated immediate treatment effect equals 445.14 for

unstandardized data while it is 26.87 for standardized data. The between-subject-variance for the estimated treatment effect on the time trend equals 0.29 for unstandardized data and 0.0011 for standardized data. Similar conclusions are obtained for the estimated between-study variances. The empirical illustration indicates that there is a large difference in estimated treatment effects, depending on whether standardized single-case data or unstandardized data are used. Because not standardizing data not measured on the same scale will flaw the results, while standardizing data that are measured on the same scale is not expected to have a systematic effect, we recommend standardizing single-case data before combining them over cases and over studies.

3.5 Conclusion and Discussion

3.5.1 General conclusion

The purpose of this study was to evaluate the standardization method proposed by Van den Noortgate and Onghena (2008). We first standardized the SSED data using this method before combining them using the three-level model.

The study indicates that the standardizing factor works reasonably well, except when a study includes a small amount of measurement observations within a subject (10 or less). In this condition, the estimate of the average treatment effects and the standard errors are underestimated which result in flawed coverage proportions.

Despite this, the results are encouraging for researchers interested in the average immediate treatment effect and the average treatment effect on the time trend, especially when including many studies (30 or more), at least 20 measurement occasions within a subject and when the studies are homogeneous (i.e., a small between-study variance). In these conditions, the average treatment effects are not biased, the mean squared error is reasonably close to zero, the coverage proportion of the .95 confidence intervals is close to the nominal level of .95. Moreover, the power of the treatment effects we tested attain the threshold level of .80 if a large number of studies (30 or more) are involved. For the treatment effect on the time trend, not only a large number of studies, but also a large number of measurements (20 or more) and homogeneous studies and cases are needed. It is important that researchers combining standardized single-subject data are aware that for other conditions substantial problems occurred.

Researchers who are interested in the variation in treatment effects over subjects and over studies should interpret the results with care. We found that estimates of the between-

subject variance are often biased, except when there are at least 40 measurement occasions within a subject. The estimation of the between-study variance is less biased, but the relative bias remains substantial if 30 studies with only 10 measurement occasions within a subject are combined. For both estimated variance components, the number of measurements has a large effect on the bias. We have to be careful with the interpretation of the bias because the value representing a substantial bias depends on the research domain and the content of the study.

The empirical illustration indicates that it is important that we standardize the data if subjects are not measured on the same scale. The estimated treatment effects differed when we use the standardized method in comparison with using the unstandardized data.

A requirement for obtaining more accurate estimated treatment effects over subjects and over studies is to use at least 20 measurement occasions per subject. Thus, we encourage single-subjects researchers to observe and measure their subjects at least 20 times.

3.5.2 Limitations and suggestions for future research

Although we tried to simulate realistic data to explore the appropriateness of the three-level model, the simulation study has the same limitation as other simulation studies, in that the conclusions are difficult to generalize to other conditions. Therefore we included conditions that are representative for the three-level analyses of single-subjects (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011). If a researcher is interested in the power when including other values for the different conditions, several tools are available. For instance, Cools, Van den Noortgate, and Onghena (2008) developed a user-friendly tool, MultiLevel Design Efficiency Using Simulation (MLdeS), that can be used for calculating the number of units needed at each level to obtain a power larger than .80. The user has to specify the model of interest (e.g. the number of levels, the number of variables at each level, the covariance structures and the parameter values) and the number of sample sizes at each level for which data will be generated, and will receive power and accuracy estimates for the parameters of interest.

Another limitation is that we assumed linear trajectories in the treatment phase, but again this is a simplification for the reality. There might be non-linear trends, and this should be a topic for future research.

There are concerns regarding the assumption that the errors in the statistical model are independent. When repeated observations are made on the same subject, there is a likelihood that the errors of the measurement associated with a score at one data point may be predictive of errors at other points in the series that follow. Subsequent measurements are more similar

than measurements farther in time: “Everything is related to everything else, but near things are more related than distant things.” (Tobler, 1970, p. 236). We did not account for this autocorrelation. Another assumption we made is that the residuals at each level are (multivariate) normally distributed. Further research is needed to investigate the performance of the approach if the normality assumption that is made in the multilevel approach is violated. A possible way to investigate this is by generating the second and third level errors from a non-normal distribution, for instance a distribution with heavier tails such as the t-distribution with a small degrees of freedom or a skewed distribution such as a χ^2-distribution. Further research is needed to investigate the consequences on the estimated treatment effect and variance components. We also assumed that residuals at each level are identically distributed, which may also not be the case. Currently, Moeyaert, Ugille, Ferron, Beretvas, and Van den Noortgate (2014a) are busy with misspecification issues in order to evaluate the robustness of the three-level approach.

To simplify the simulation model, we did not account for a possible dependence between different regression coefficients, that can be accounted for in a multilevel analysis by estimating the covariances at the various levels. For instance, it seems plausible that the treatment effects, β_{2jk} and β_{3jk}, are smaller for subjects with an already high baseline level, β_{0jk}. Further research is needed to explore level-2 and level-3 covariance matrix misspecification.

We used a multiple-baseline across participants design but we did not take into account confounding extraneous events that could have a simultaneous effect on all participants. In current research, we developed a method to model these extraneous events (Moeyaert et al., 2013c). Beside this, there are other designs like alternating designs and reversal designs that need further exploration.

Finally, parameters were estimated and tested using REML, which is based on large sample theory. The results indicate that the variance components at all levels were biased. We are currently doing more research, evaluating the approach in more complex conditions, exploring alternative methods for estimating the parameters and evaluating ways to avoid biased parameter and corresponding standard error estimates for the analysis of standardized SSED data.

This research indicates that the three-level approach is appropriate to combine standardized single-subject data as long as the studies are quite homogeneous, there are a lot of measurements per subject and a lot of studies are combined.

Chapter 4|

Modeling External Events in the Three-Level Analysis of Multiple-Baseline Across Participants Design[3]

Abstract

In this study, we focus on a three-level meta-analysis for combining data from studies using multiple-baseline across participants designs. A complicating factor in such designs is that results might be biased if the dependent variable is affected by not explicitly modeled external events, such as the illness of a teacher, an exciting class activity, or the presence of a foreign observer. In multiple-baseline designs, external effects can become apparent if they simultaneously have an effect on the outcome score(s) of the participants within a study. This study presents a method to adjust the three-level model for external events and evaluates the appropriateness of the modified model. Therefore we use a simulation study, and we illustrate the new approach with real datasets.

The results indicate that ignoring an external event effect results in biased estimates of the treatment effects, especially when there is only a small number of studies and measurement occasions involved. The mean squared error, as well as the standard error and coverage proportion of the effect estimates are improved with the modified model. Moreover, the adjusted model results in less biased variance estimates. If there is no external event effect, we find no differences in results between the modified and unmodified models.

Keywords: multiple-baseline across participants, three-level meta-analysis, effect sizes, external event effect

[3] This chapter has been published as Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S.N.,& Van den Noortgate, W. (2013c). Modeling external events in the three-level analysis of multiple-baseline across participants designs: A simulation study. *Behavior Research Methods, 45*, 547-559. doi: 10.1080/00273171.2013.816621

4.1 Introduction

4.1.1 Multiple-baseline design

A multiple-baseline design (MBD) is one of the variants of Single-Subject Experimental Designs (SSEDs). SSED researchers observe and measure a participant or case repeatedly over time. Observations are obtained during at least one baseline phase (when no intervention is present) and at least one treatment phase (when an intervention is present). By comparing scores from both kinds of phases, SSED researchers can assess whether the outcome scores on the dependent variable changed for instance in level or in slope when the treatment was present (Onghena & Edgington, 2005).

In an MBD, an AB phase design (with one baseline phase, A, and one treatment phase, B) is implemented simultaneously to different participants, behaviors or settings (Barlow & Hersen, 1984; Ferron & Scott, 2005; Onghena, 2005; Onghena & Edgington, 2005). MBDs are popular amongst the SSEDs (Shadish & Sullivan, 2011) because the intervention is introduced sequentially over the participants (or settings or behaviors), which entails the advantage that researchers can more easily disentangle effects of the intervention and effects of some external events, such as the illness of a teacher, an exciting class activity, the presence of a foreign observer, and a teacher intern (Baer et al., 1968; Barlow & Hersen, 1984; Kinugasa et al., 2004; Koehler & Levin, 2000). This is because if an external event occurs at certain points in time, then the outcome scores for all participants in that study might be simultaneously influenced. Figure 4.1 gives a graphical presentation of possible consequences for the occurrence of an external event in a multiple-baseline across participants design. In Figure 4.1a, the external event has a constant effect on the dependent variable on subsequent measurements, for instance the teacher is ill during subsequent days, or there is a foreign observer during some measurement occasions. Figure 4.1b illustrates a gradually fading away external event effect. For instance, the influence of a teacher intern on the behavior of the students may be reduced over time.

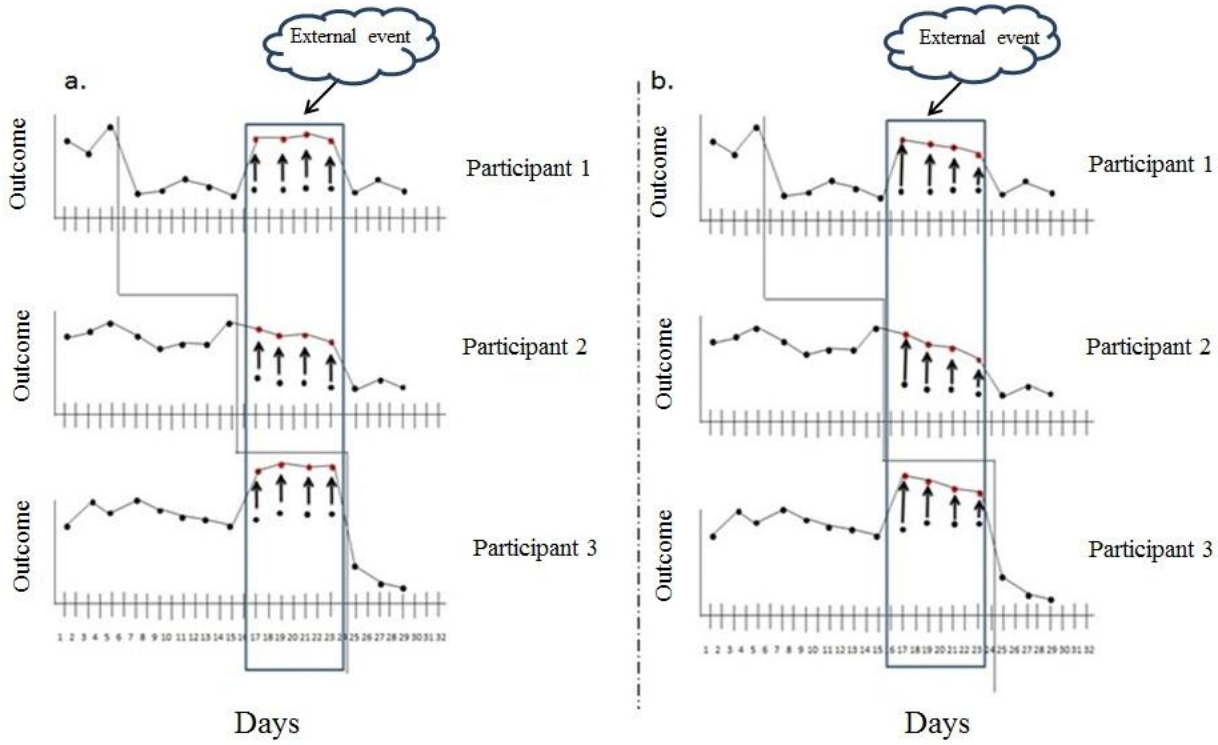

Figure 4.1. Graphical display of a constant external event effect (a) and a gradually fading away external event effect (b) affecting the score on 4 subsequent moments (day 17, day 19, day 21 and day 23) for a MBD across three participants with the treatment starting on day 6, day 16 and day 24 respectively.

4.1.2 Multilevel meta-analysis

Van den Noortgate and Onghena (2003b) proposed the use of multilevel models to synthesize data from multiple SSED studies, allowing investigation of the generalizability of the results, and exploration of potential moderating effects. In previous research evaluating this multilevel meta-analysis of MBD data (Ferron et al., 2009; Ferron et al., 2010; Moeyaert et al., 2013a, 2013b; Owens & Ferron, 2012), the data were typically simulated with a treatment effect and random noise only. Potential confounding events that could have a simultaneous effect on all participants within a study were not taken into account. In this study we evaluate the performance of the basic three-level model when there are effects of external events, as well as of an extension of the model that tries to account for potential event effects. In the following, we first present the basic model and a possible extension to account for external events. Next, we evaluate the performance of both models, by means of a simulation study and an analysis of real data.

A meta-analysis combines the results of several studies addressing the same research question (Cooper, 2010; Glass, 1976). Study results are typically first converted to a common standardized effect size before meta-analyzing them. The effect sizes may be reported in the primary studies or can be calculated afterwards, using reported summary and/or test statistics.

One possible way to calculate effect sizes when using SSEDs, is to analyze the data using regression models, and to use the regression coefficients as effect sizes. A regression model of interest here is the one proposed by Center et al. (1985-1986):

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 T'_i D_i + e_i \text{ with } e_i \sim N(0, \sigma_e^2) \quad (4.1)$$

The score of the dependent variable on measurement occasion *i* (Y_i) depends on a dummy coded variable (D_i) indicating whether the measurement occasion *i* belongs to the baseline phase (D_i = 0) or the treatment phase (D_i = 1), a time-related variable T_i, that equals 1 on the first measurement occasion of the baseline phase, and an interaction term between the centered time-indicator and the dummy variable, $T'_i D_i$, where T'_i is centered such that T'_i equals 0 on the first measurement occasion of the treatment phase. β_0 indicates the expected baseline level, β_1 is the linear trend during the baseline, β_2 refers to the immediate treatment effect and β_3 to the effect of the treatment on the time trend.

Van den Noortgate and Onghena (2003b) proposed using the ordinary least squares estimates for β_2 and β_3 from Equation 4.1 as effect sizes in the three-level meta-analysis. At the first level the estimated effect sizes of the immediate treatment effect, b_{2jk}, and the treatment effect on the time trend, b_{3jk}, for participant *j* from study *k* are equal to the unknown population effects sizes, β_{2jk} and β_{3jk} respectively, plus random deviations, r_{2jk} and r_{3jk}, that are assumed to be normally distributed with a mean of zero:

$$\begin{aligned} b_{2jk} &= \beta_{2jk} + r_{2jk} \quad \text{with} \quad r_{2jk} \sim N\left(0, \sigma^2_{r_{2jk}}\right) \\ b_{3jk} &= \beta_{3jk} + r_{3jk} \quad \text{with} \quad r_{2jk} \sim N\left(0, \sigma^2_{r_{3jk}}\right) \end{aligned} \quad (4.2)$$

The sampling variances of the observed effects, $\sigma^2_{r_{2jk}}$ and $\sigma^2_{r_{3jk}}$ are the squared standard errors that are typically reported by default when performing a regression analysis. These variances depend to a large extent on the number of observations and the variance of these observations, and therefore can be participant- and study-specific.

At the second level, the population effect sizes β_{2jk} and β_{3jk} from Equation 4.2 can be modeled as varying over participants around the study-specific mean effect, θ_{20k} and θ_{30k} (Equation 4.3).

$$\begin{aligned} \beta_{2jk} &= \theta_{20k} + u_{2jk} \quad \text{with} \quad u_{2jk} \sim N\left(0, \sigma^2_{u_{2jk}}\right) \\ \beta_{3jk} &= \theta_{30k} + u_{3jk} \quad \text{with} \quad u_{3jk} \sim N\left(0, \sigma^2_{u_{3jk}}\right) \end{aligned} \tag{4.3}$$

The population effects for studies can vary between studies (third level, Equation 4.4).

$$\begin{aligned} \theta_{20k} &= \gamma_{200} + v_{20k} \quad \text{with} \quad v_{20k} \sim N\left(0, \sigma^2_{v_{20k}}\right) \\ \theta_{30k} &= \gamma_{300} + v_{30k} \quad \text{with} \quad v_{30k} \sim N\left(0, \sigma^2_{v_{30k}}\right) \end{aligned} \tag{4.4}$$

The model parameters that we are typically interested in when using a multilevel model are the fixed effects regression coefficients (i.e., γ_{200}, referring to the average immediate treatment effect over participants and studies and γ_{300}, referring to the average treatment effect on the linear trend over participants and studies in Equations 4.4), the variances (i.e., $\sigma^2_{v_{20k}}$, referring to the between-study variance for the estimated immediate treatment effect, $\sigma^2_{v_{30k}}$, indicating the between-study variance for the estimated treatment effect on the time trend, $\sigma^2_{u_{2jk}}$, the between-case variance for the estimated immediate treatment effect and $\sigma^2_{u_{3jk}}$, referring to the between-case variance of the estimated treatment effect on the time trend).

4.1.3 Correcting effect sizes for external events

External events in a multiple-baseline across participants design, can have an effect on the outcome score(s) of all participants within a study. These external event effects are common in SSEDs, because practitioners often implement these designs in their everyday setting (for example in the home, - school, - etc.), where they cannot control for outside experimental factors (Christ, 2007; Kratochwill et al., 2010; Shadish et al., 2002). If we do not model these external events, the results might be biased. For instance, suppose a researcher is interested in a change in challenging behavior and staggered the beginning of the treatment across three participants. The three participants receive the treatment at day 6, day 16 and day 24, respectively (see Figure 4.1) and are observed every 2 days. On day 17, 19, 21 and 23, the teacher is ill and as a consequence a substitute teacher takes their place and the participants exhibit more challenging behavior. In this situation, the estimated treatment effect for participant 1 and 2 will be smaller and the estimated treatment effect for participant

3 will be larger, and therefore differences between participants in the treatment effects are also likely to be overestimated, unless we correct the effect sizes for possible external events.

A possible way to calculate effect sizes corrected for an external event in an SSED is by estimating effect sizes for participants per study, by performing a regression analysis with a model including possible event effects, and assuming that external events simultaneously affect all participants in a study. Thereafter, the corrected effect sizes can be combined over studies in the three-level meta-analysis.

For the first step, we propose to use an extension of the Center et al. (1985-1986) model, including dummy variables for measurement occasions:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}D_{ij} + \beta_{3j}D_{ij}T'_{ij} + \sum_{m=2}^{I-1} \beta_{(m+2)}M_{mi} + e_{ij} \text{ with } e_{ij} \sim N(0, \sigma_e^2) \quad (4.5)$$

The score on the dependent variable *Y* on measurement occasion *i* (=1, 2, …, *I*) from participant *j (*=1, 2, …, *J)* is modeled as a linear function of the dummy coded variable (D_{ij}) indicating whether the measurement occasion *i* from participant *j* belongs to the baseline phase (D_{ij} = 0) or the treatment phase (D_{ij} = 1), a time-related variable T_{ij}, that equals 1 at the start of the baseline phase, an interaction term between the dummy variable indicating the phase and the time-indicator centered around its value at the start of the treatment phase, $D_{ij}T'_{ij}$, and finally dummy coded variables indicating the moment (M_{mi} = 1 if $m = i$, zero otherwise). By including the effects of individual moments, coefficients β_{2j} and β_{3j} can be interpreted as the treatment effects, corrected for possible external events.

We do not include a dummy variable for one measurement moment in the baseline phase and one measurement moment in the treatment phase. This is to ensure that the model is identified: if we would include these parameters as well, an increase in the effects for each moment in the baseline phase, could be compensated by a decrease of the intercept, illustrating that without constraining these parameters, there would be an infinite number of equivalent solutions. For our study we select the first and last moment as the times to set the moment effects to zero, but different moments could be chosen if we suspected a moment effect during one of these times.

While the baseline level and slope (β_{0j} and β_{1j}) and both treatment effects (β_{2j} and β_{3j}) are participant-specific, the moment effects are assumed to be the same for all participants from the same study, and therefore have to be estimated for each study using all data from that study. To this end, we propose to extend Equation 4.5 by including a set of

dummy participant indicators. For two participants, using dummy participant indicators P_1 and P_2 respectively, this results in Equation 4.6:

$$Y_{ij} = \beta_{01}P_{1j} + \beta_{02}P_{2j} + \beta_{11}T_{i1}P_{1j} + \beta_{12}T_{i2}P_{2j} + \beta_{21}D_{i1}P_{1j} + \beta_{22}D_{i2}P_{2j} + \beta_{31}D_{i1}T'_{i1}P_{1j} + \beta_{32}D_{i2}T'_{i2}P_{2j} + \sum_{m=2}^{I-1}\beta_{(m+2)}M_{mi} + e_{ij} \quad (4.6)$$

$$\text{with } e_{ij} \sim N(0, \sigma_e^2)$$

After using Equation 4.6 for each study to estimate the corrected effect sizes (β_{2j} and β_{3j}) for each participant, we can use the three-level meta-analysis (see Equation 4.2 - 4.4) to combine the corrected effect size estimates from multiple participants. In principle we could also use a two-level model per study to estimate the participant-specific effects, but given the typically very small number of participants per study, using a multilevel model might not be recommended.

4.2 A Simulation Study

4.2.1 Simulating three-level data

To evaluate the performance of the basic model and its extension, we performed a simulation study. We simulated raw data using a three-level model. At level one, we used the following model:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}T'_{ijk}D_{ijk} + e_{ijk} \text{ with } e_{ijk} \sim N(0, \sigma_e^2) \quad (4.7)$$

with measurement occasions nested within participants, which form the units at level two:

$$\begin{cases} \beta_{0jk} = \theta_{00k} + u_{0jk} \\ \beta_{1jk} = \theta_{10k} + u_{1jk} \\ \beta_{2jk} = \theta_{20k} + u_{2jk} \\ \beta_{3jk} = \theta_{30k} + u_{3jk} \end{cases} \text{ with } \begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{2jk} \\ u_{3jk} \end{bmatrix} \sim N(0, \Sigma_u) \quad (4.8)$$

The participants are in turn clustered within studies at the third level:

$$\begin{cases} \theta_{00k} = \gamma_{000} + v_{00k} \\ \theta_{10k} = \gamma_{100} + v_{10k} \\ \theta_{20k} = \gamma_{200} + v_{20k} \\ \theta_{30k} = \gamma_{300} + v_{30k} \end{cases} \text{ with } \begin{bmatrix} v_{00k} \\ v_{10k} \\ v_{20k} \\ v_{30k} \end{bmatrix} \sim N(0, \Sigma_v) \quad (4.9)$$

4.2.3 Varying parameter

Based on a thorough overview of 809 SSED studies, Shadish and Sullivan (2011) enumerated some parameters that characterize SSEDs. Based on their results and our re-analyses of meta-analyses of SSEDs (Alen et al., 2009; Denis et al., 2011; Ferron et al., 2010; Kokina & Kern, 2010; Shadish & Sullivan, 2011; Shogren et al., 2004; Wang et al., 2011), we decided to vary the following parameters that can have a significant influence on the quality of model estimation:

- γ_{200}, represents the immediate treatment effect on the outcome and had values 0 (no effect) or 2.
- The treatment effect on the time trend, defined by γ_{300}, was varied to have values 0 (no effect) or 0.2.
- The regression coefficients of the baseline, γ_{000} and γ_{100}, did not vary and were set at 0, because the interest is in the average treatment effects (i.e. the immediate treatment effect and the treatment effect on the time trend).
- The number of simulated participants, *J*, equaled 4 or 7.
- The number of measurements within a participant, *I*, was 15 or 30. We chose to keep *I* constant for all participants within the same study.
- The number of studies, *K*, was 10 or 30.
- The between case-covariance matrix: covariances between pairs of regression coefficients were set to zero. Therefore, Σ_u is a diagonal matrix. $\Sigma_u = \text{diag}(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3})$ = diag(2, 0.2, 2, 0.2) or $\Sigma_u = \text{diag}(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3})$ = diag(0.5, 0.05, 0.5, 0.05).
- The between study-covariance matrix: covariances between pairs of regression coefficients were set to zero. Therefore, Σ_v is a diagonal matrix. $\Sigma_v = \text{diag}(\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v}, \sigma^2_{v_3})$ = diag(2, 0.2, 2, 0.2) or $\Sigma_v = \text{diag}(\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3})$ = diag(0.5, 0.05, 0.5, 0.05).
- The moment of introducing a treatment effect was staggered across participants within a study (see Table 4.1), depending on the number of measurements.

Table 4.1

Time of Introducing the Treatment

	Start of intervention						
I	articipant 1	participant 2	participant 3	participant 4	participant 5	participant 6	participant 7
15	5	6	7	8	9	11	13
30	5	8	11	14	17	20	23

In a first scenario, a constant external event was added to influence four subsequent scores of all the participants within a study (as in Figure 4.1a). The moment was randomly generated from a uniform distribution for each study separately. Because we did not include a moment effect for the first and the last moment to make the model identified, the external event effect did not occur on these moments. The external event effect was 0 or 2, representing a null and a large external event effect, respectively.

In a second scenario, the effect of the external event effect was added that fades away gradually (see Figure 4.1b) for all the participants within a study. The effect across four time points was respectively 3.5, 2.5, 1.5, 0.5. or 0, so that on average the average effect was the same as in the first scenario. The start of the event effect was generated completely at random from a uniform distribution for each study separately, so that the external event effect did not occur on the first or last measurement occasion. Data were generated using SAS 9.3.

4.2.4 Analysis

We had a total of 2^9 (= 512) experimental conditions. We simulated 400 replications of each condition, resulting in 204,800 datasets to analyze. We analyzed the data twice, and compared the results. First we combined the uncorrected effect sizes in the three-level meta-analysis. Next, we analyzed the three-level data by estimating the corrected effect sizes, β_{2j} and β_{3j}, using the regression analysis per study (see Equation 4.6) before combining them in the three-level meta-analysis (see Equation 4.2 - 4.4).

In the two approaches we used the SAS PROC MIXED (Littell et al., 2006) procedure to estimate the participant-specific effect sizes, β_{2jk} and β_{3jk}. In the first approach the effect sizes were uncorrected for the external event effect, whereas the effect sizes in the second approach were corrected.

SAS PROC MIXED was also used for the three-level meta-analysis. The Satterthwaite approach to estimate the degrees of freedom method was applied because this method provides more accurate confidence intervals for estimates of the average treatment effect for two-level analyses of multiple-baseline data (Ferron et al., 2009).

In order to evaluate the appropriateness of both models, uncorrected and corrected for external events, we calculated the deviations of the estimated immediate treatment effect, $\hat{\gamma}_{200}$, from its population value, γ_{200}, and the deviations of the estimated treatment effect on the time trend, $\hat{\gamma}_{300}$, from its population value, γ_{300}. The mean deviation gives us an idea of the bias. Next, we calculated the mean squared deviation (the *Mean Squared Error*, *MSE*) which gives information about the variance of both estimated treatment effects ($\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$) around the corresponding population effect (γ_{200} and γ_{300}). Furthermore we discuss the standard error and the 95% confidence interval coverage proportion (*CP*) of the estimated immediate treatment effect and the treatment effect on the time trend. We also evaluate the bias of the point estimates of the between-study and between-case variance.

We used ANOVAs to evaluate whether there were significant effects ($\alpha = .01$) of each model type (e.g. model using effect sizes corrected versus uncorrected for external event effects) and of the simulation design parameters (γ_{200}, γ_{300}, *K*, *I*, *J*, $\sigma^2_{u_2}$, $\sigma^2_{v_2}$) on the bias, *MSE*, the standard error and the *CP*.

4.3 Results of the Simulation Study

We present the results in two sections. In the first section we discuss the constant external effect over four subsequent measurement occasions. The second section considers the case where the external effect gradually fades away over four subsequent measurements. Each section presents the results of the three-level analysis of uncorrected and corrected effect sizes.

When there is no external event effect, the results of the three-level meta-analysis (i.e., bias in the fixed effects, *MSE* of the fixed effects, estimated standard errors of the fixed effects, *CP* for the fixed effects, and bias in the variance components) were found to be independent of the model type (corrected or not corrected for external events).

We found no significant bias for $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ when using the corrected or uncorrected model. Therefore we only discuss the results of the analyses of the data including external event effects conditions.

4.3.1 Constant external event over four subsequent measurement occasions

4.3.1.1 Average treatment effect

4.3.1.1.1 Bias and mean squared error

When we estimate γ_{200} and the effect sizes are uncorrected, the estimated treatment effect is on average significantly larger than the population value (γ_{200} = 0 or 2). Over all conditions, the bias equals 0.032, $t(51199) = 17.32$, $p < .0001$, whereas there is no significant bias for the corrected effect sizes; -0.0015, $t(51199) = -0.96$, $p = .34$. Table 4.2 presents the bias estimates for $\hat{\gamma}_{200}$, when γ_{200}= 2 and γ_{300} = 0.2.

Similar results are obtained for $\hat{\gamma}_{300}$. The bias is significantly negative for the uncorrected effect sizes and equals -0.20, $t(51199) = -255.27$, $p < .0001$, whereas the bias is not significant for the corrected effect sizes, $t(51199) = -0.00020$, $p = .79$. Moreover, an analysis of variance on the deviations reveals a significant difference between the two different models, both for $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. $F(1, 102398) = 192.06$, $p < .0001$ for $\hat{\gamma}_{200}$ and $F(1, 102398) = 33695.1$, $p < .0001$ for $\hat{\gamma}_{300}$. The differences are largest when there is a small number of measurement occasions ($I = 15$) and studies ($K = 10$). In the following condition the largest difference was identified: γ_{200} = 2, γ_{300} = 0, K = 10, I = 15, J = 4, $\sigma^2_{u_2}$ = 0.5 and $\sigma^2_{v_2}$ = 2 (with a difference of 0.23).

Table 4.2

The Bias of $\hat{\gamma}_{200}$; for γ_{200} = 2 and *γ_{300} = 0.2 Conditions for the Constant External Event Effect over 4 Subsequent Measurement Occasions*

			Corrected				Unorrected			
			I = 15		I = 30		I = 15		I = 30	
K	J	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2
10	4	0.5	-0.003	0.007	0.025	-0.036	0.213	0.208	-0.027	0.027
		2	0.015	0.002	-0.017	0.014	0.129	0.196	0.012	0.035
	7	0.5	-0.026	-0.057	0.024	0.005	-0.093	-0.058	-0.019	-0.074
		2	-0.028	-0.015	-0.011	-0.003	-0.099	-0.060	-0.016	-0.026
30	4	0.5	0.009	0.028	0.004	-0.005	0.219	0.185	-0.008	0.013
		2	0.018	0.021	0.004	-0.011	0.210	0.222	0.008	0.035
	7	0.5	0.023	0.005	0.002	-0.009	-0.075	-0.105	-0.004	-0.016
		2	0.001	0.026	-0.006	-0.012	-0.077	-0.088	-0.003	0.006

Note. Corrected and Uncorrected refer respectively to corrected effect size and uncorrected effect size for external event effects.

Similar to the bias, the Mean Squared Error (*MSE)* of the estimated treatment effect depends significantly on the model type; using an analysis of variance on the squared deviations, $F(1, 102398) = 882.77$, $p < .0001$ for $\hat{\gamma}_{200}$ and $F(1, 102398) = 7076.91$, $p < .0001$ for $\hat{\gamma}_{300}$. When using the corrected model, the *MSE* for respectively $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ equals: 0.12 and 0.028, whereas it is 0.18 and 0.070, respectively, for the uncorrected effect sizes. Differences between both models are larger if the number of observations and the number of studies are small (see Table 4.3 for $\hat{\gamma}_{200}$, similar results are obtained for $\hat{\gamma}_{300}$). So especially in these conditions the modified model is recommended.

Table 4.3

The MSE of $\hat{\gamma}_{200}$; for γ_{200} = 2 and γ_{300} = 0.2 Conditions for the Constant External Event Effect over 4 Subsequent Measurement Occasions

			Corrected				Unorrected			
			I = 15		*I* = 30		*I* = 15		*I* = 30	
K	*J*	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 0.5	$\sigma^2_{v_2}$ = 2
10	4	0.5	0.17	0.28	0.11	0.26	0.32	0.43	0.14	0.25
		2	0.20	0.32	0.14	0.28	0.31	0.49	0.16	0.36
	7	0.5	0.09	0.24	0.07	0.23	0.18	0.31	0.09	0.22
		2	0.11	0.26	0.09	0.24	0.20	0.31	0.09	0.28
30	4	0.5	0.06	0.10	0.04	0.09	0.14	0.19	0.04	0.10
		2	0.06	0.11	0.04	0.09	0.15	0.20	0.05	0.12
	7	0.5	0.03	0.07	0.03	0.08	0.06	0.10	0.03	0.09
		2	0.04	0.08	0.03	0.08	0.07	0.10	0.04	0.08

Note. Corrected and Uncorrected refer respectively to corrected effect size and uncorrected effect size for external event effects.

4.3.1.1.2 Estimates of the standard errors

In order to evaluate inferences regarding the treatment effects, we constructed confidence intervals around the estimated treatment effects, $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. Therefore we needed to estimate the standard errors of the estimated treatment effects. Because we obtained 400 estimates of the effects in each condition, the standard deviations of the effect estimates can be regarded as a relatively good estimate of the standard deviation of the sampling distribution, and can therefore be used as a criterion to evaluate the standard error. We looked at the relative standard error biases which are the differences between the median standard error estimates and the standard deviation of the estimates of the effect divided by the standard deviation of the estimates of $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. The relative differences are negative for $\hat{\gamma}_{200}$ which means that the median standard error estimates are smaller than expected. For $\hat{\gamma}_{300}$, these differences are positive, referring to median standard error estimates larger than expected. The relative standard error biases for both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ are on average larger

across the conditions for the uncorrected effect sizes in comparison with the corrected effect sizes. For $\hat{\gamma}_{200}$, the average relative standard error biases equal -1.8% and -2.0% for the corrected and uncorrected model respectively. The average relative standard error biases difference for $\hat{\gamma}_{300}$ for the uncorrected model is 2% whereas it is substantial (more than 10%; Hoogland & Boomsma; 1998) for the uncorrected model; 25.7%. So the difference between the model type becomes more apparent when estimating γ_{300}, $F(1, 254) = 38.9$, $p < .0001$. The conditions with the largest relative standard error bias when using the uncorrected model for $\hat{\gamma}_{300}$ tended to coincidence with the conditions where 30 studies, an immediate treatment effect of 2 and a treatment effect on the time trend of 0.2 were involve with the bias mounting to 107% in the condition where γ_{200}= 2, γ_{300} = 0.2, K = 30, J = 7, I = 30, $\sigma^2_{v_2}$= 0.5 and $\sigma^2_{u_2}$= 0.5.

4.3.1.1.3 Coverage proportion

We estimated the coverage proportion (*CP*) of the 95% confidence intervals which allows us to evaluate the interval estimates of $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. The confidence intervals were estimated by using the standard errors and the Satterthwaite estimated degrees of freedom. The *CP* of these confidence intervals was estimated for each of the combinations. A positive significant difference between the corrected model and the uncorrected model in the *CP* is found for $\hat{\gamma}_{200}$, $F(1, 254) = 27.56$, $p < .0001$ (see Table 4.4). Also for $\hat{\gamma}_{300}$, the mean *CP* depends significantly on the model type, $F(1, 254) = 20.96$, $p < .0001$ (see Table 4.4). The conditions with a *CP* less than .93 all have 15 measurements in common and occur when the effect sizes are uncorrected, for both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. Moreover, for $\hat{\gamma}_{300}$, the *CP* is not only too small when I = 15 and K = 30, but also too large when I = 30 (values for the *CP* range from .99 to 1.00). When the effect sizes are uncorrected, the *CP* is well estimated when I = 30 for $\hat{\gamma}_{200}$ and I = 15 and K = 10 for $\hat{\gamma}_{300}$. The difference in *CP* for $\hat{\gamma}_{200}$ is largest when there are only a small number of measurements ($I = 15$) and a large number of studies ($K = 30$).

Table 4.4

The Coverage Proportion of $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$; for γ_{200} = 2, γ_{300} = 0.2 and $\sigma^2_{u_2}$ = 2 Conditions for the Constant External Event Effect over 4 Subsequent Measurement Occasions

			$\hat{\gamma}_{200}$				$\hat{\gamma}_{300}$			
			Corrected		Uncorrected		Corrected		Uncorrected	
K	*J*	$\sigma^2_{v_2}$	*I* = 15	*I* = 30	*I* = 15	*I* = 30	*I* = 15	*I* = 30	*I* = 15	*I* = 30
10	4	0.5	.96	.96	.96	.96	.94	**1.00**	.97	**1.00**
		2	.95	.95	**.92**	.95	.96	**.98**	.93	**1.00**
	7	0.5	.96	.95	.94	.97	**.99**	**1.00**	.97	**1.00**
		2	.97	.96	.95	.95	.96	.97	**.84**	**.99**
30	4	0.5	.97	.96	**.89**	.97	.97	**1.00**	**.90**	**1.00**
		2	.97	.96	**.91**	.94	.96	**.98**	**.49**	**1.00**
	7	0.5	.94	.94	**.92**	.96	**.98**	**1.00**	.93	**1.00**
		2	.96	.95	.96	.96	.96	.97	**.26**	.96

Note. Values smaller than .93 and larger than .97 appear in bold. Corrected and Uncorrected refer respectively to corrected effect size and uncorrected effect size for external event effects.

4.3.1.2 Variance components

In the three-level analyses, the between-study and between-case variances were estimated for both the immediate treatment effect and the treatment effect on the trend. Because variance estimates are expected to be positively skewed, due to truncation of negative estimates to zero, we calculated the median (relative) deviation of the estimates from the population value, rather than the mean (relative) deviation, to evaluate the (relative) bias in the estimates. We only discuss the between-case variance and the between-study variance of the immediate treatment effect ($\sigma^2_{u_2}$ and $\sigma^2_{v_2}$), because similar results are obtained for the treatment effect on the time trend ($\sigma^2_{u_3}$ and $\sigma^2_{v_3}$). The bias of the estimated between-study variance and the estimated between-case variance of the immediate effect is larger when there are only 10 studies and 15 measurement occasions involved. The conditions with the largest relative bias all had 15 measurements, 4 participants and a small between-study variance ($\sigma^2_{v_2}$ = 0.5) in common. If the effect sizes are corrected and we estimate the between-study variance of the immediate treatment effect, we find relative parameter bias values across conditions ranging from 17% to 55%, while the relative bias goes up to a value of 313% when the effect sizes are uncorrected. Similar results are found for $\hat{\sigma}^2_{u_2}$, where the relative bias in a condition is maximum 119% for the corrected effect sizes and 326% for the uncorrected effect sizes (see Table 4.5). Overall, the adjusted model results in less biased variance estimates.

Table 4.5

Median of Relative Deviation of the Variance Estimates of γ_{200}, for γ_{200} = 2 *and* γ_{300} = 0.2 *Conditions for the Constant External Event Effect over 4 Subsequent Measurement Occasions*

			$\hat{\sigma}^2_{v_2}$								$\hat{\sigma}^2_{u_2}$							
			Corrected				Uncorrected				Corrected				Uncorrected			
			$K = 10$		$K = 30$		$K = 10$		$K = 30$		$K = 10$		$K = 30$		$K = 10$		$K = 30$	
I	J	$\sigma^2_{v_2}$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$	$\sigma^2_{u_2}=0.5$	$\sigma^2_{u_2}=2$
15	4	0.5	0.47	0.61	0.55	0.58	2.70	3.09	3.11	2.91	1.24	0.24	1.19	0.27	3.01	0.72	3.13	0.79
		2	0.05	-0.03	0.13	0.12	0.71	0.57	0.74	0.65	1.16	0.20	1.19	0.31	3.11	0.75	3.26	0.80
	7	0.5	-0.07	-0.11	0.09	0.06	0.99	0.99	1.10	0.98	1.15	0.25	1.20	0.29	2.87	0.71	2.85	0.70
		2	-0.02	-0.08	0.01	-0.02	0.15	0.16	0.29	0.28	1.18	0.27	1.20	0.28	2.79	0.63	2.93	0.73
30	4	0.5	-0.12	-0.09	-0.02	0.05	0.33	0.26	0.45	0.48	0.83	0.21	0.80	0.17	0.61	0.13	0.81	0.20
		2	-0.08	-0.06	-0.01	-0.04	0.14	0.10	0.08	0.10	0.81	0.16	0.85	0.21	0.59	0.10	0.80	0.20
	7	0.5	-0.17	-0.12	0.02	-0.03	-0.08	-0.07	0.04	0.04	-0.12	-0.06	-0.1	-0.03	0.94	0.19	0.89	0.22
		2	-0.11	-0.12	-0.01	-0.04	-0.07	-0.02	0.00	0.00	-0.11	-0.05	-0.12	-0.01	0.91	0.23	0.92	0.22

Note. Corrected and Uncorrected refer respectively to corrected effect size and uncorrected effect size for external event effects.

4.3.2 External event fades away gradually over four subsequent measurement occasions

4.3.2.1 Average treatment effect.

4.3.2.1.1 Bias and mean squared error

The bias of $\hat{\gamma}_{200}$ for uncorrected and corrected effect sizes is respectively -.0073, $t(51199) = -418$, $p < .001$ and 0.00057, $t(51199) = 38$, $p = .74$. This means that there is a significant negative bias for the uncorrected effect sizes, whereas this is not the case for the corrected effect sizes and the models differ significantly, $F(1, 102398) = 0.009$, $p = .77$. The bias for $\hat{\gamma}_{300}$ depends largely on the model type, $F(1, 102398) = 30476.1$, $p < .0001$. The bias for the uncorrected effect sizes is significant: -0.19, $t(51199) = -246.23$, $p < .0001$, whereas this is not the case for the corrected: 0.000179, $t(51199) = 0.24$, $p = .81$. For both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$, the difference is largest when there are a small number of measurements ($I = 15$) involved.

For both estimated treatment effects, the *MSE*'s are larger for the uncorrected effect sizes in comparison to the corrected effect sizes (see Table 4.6). For both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$, the model type has a significant influence on the *MSE*, $F(1, 102398) = 724.69$, $p = < .0001$ for $\hat{\gamma}_{200}$ and for $\hat{\gamma}_{300}$, $F(1, 102398) = 5431.15$, $p < .0001$. For both estimated treatment effects, the *MSE* is large when the studies are heterogeneous ($\sigma^2_{v_2} = 2$) and a small number of measurement occasions ($I = 15$) and studies ($K = 10$) are used. The difference between the models is largest when a small number of measurements are used.

Table 4.6

The MSE of $\hat{\gamma}_{200}$, and $\hat{\gamma}_{300}$; for $\gamma_{200} = 2$, $\gamma_{300} = 0.2$ and $\sigma^2_{u_2} = 0.5$ Conditions for the External Event Effect Fading away Gradually over 4 Subsequent Measurement Occasions

			$\hat{\gamma}_{200}$				$\hat{\gamma}_{300}$			
			Corrected		Uncorrected		Corrected		Uncorrected	
K	*J*	$\sigma^2_{v_2}$	*I* = 15	*I* = 30	*I* = 15	*I* = 30	*I* = 15	*I* = 30	*I* = 15	*I* = 30
10	4	0.5	0.14	0.11	0.34	0.12	0.09	0.01	0.12	0.01
		2	0.31	0.24	0.47	0.27	0.13	0.03	0.14	0.03
	7	0.5	0.09	0.06	0.18	0.09	0.01	0.01	0.10	0.01
		2	0.22	0.23	0.32	0.22	0.03	0.02	0.12	0.02
30	4	0.5	0.05	0.03	0.11	0.04	0.04	0.004	0.11	0.01
		2	0.11	0.09	0.17	0.09	0.04	0.01	0.12	0.01
	7	0.5	0.03	0.02	0.06	0.02	0.004	0.002	0.09	0.01
		2	0.08	0.08	0.10	0.07	0.01	0.01	0.1	0.01

Note. Corrected and Uncorrected refer respectively to corrected effect size and uncorrected effect size for external event effects.

4.3.2.1.2 Estimates of the standard errors

The difference between the average relative bias in the standard errors of the uncorrected effect sizes equals 0.02 for both uncorrected and corrected effect sizes when estimating γ_{200}.

Similar to the constant external event effect results, the difference between the average relative bias in the standard errors of the uncorrected effect sizes (M = 39.3) and corrected effect sizes (M = 0.06) for $\hat{\gamma}_{300}$ is larger and statistically significant, $F(1, 254) = 129.66$, $p = .0001$, see Table 4.7. The difference in results due to the model type is more obvious if there are a small number of studies involved ($K = 10$).

Table 4.7

Difference Between the Median of the Standard Error Estimates and the Standard Deviation of $\hat{\gamma}_{300}$; for γ_{200} = 2 , γ_{300} = 0.2 and $\sigma^2_{u_3}$= 0.05 for the External Event Effect Fading away Gradually over 4 Subsequent Measurement Occasions

			Corrected		Uncorrected	
K	*J*	$\sigma^2_{v_3}$	*I* = 15	*I* = 30	*I* = 15	*I* = 30
10	4	0.05	0.01	0.037	0.076	0.103
		0.2	-0.031	-0.002	0.029	0.05
	7	0.05	0.004	0.035	0.069	0.068
		0.2	-0.001	-0.007	0.01	0.012
30	4	0.05	-0.018	0.022	0.039	0.061
		0.2	-0.009	0.0003	0.022	0.024
	7	0.05	0.002	0.021	0.038	0.04
		0.2	-0.002	0.001	0.004	0.003

Note. Corrected and Uncorrected refer respectively to corrected effect size and uncorrected effect size for external event effects.

4.3.2.1.3 Coverage proportion

Similar to the *CP* for the constant external event effect, the mean *CP* for the uncorrected and corrected effect sizes for the estimate of the immediate treatment effect differ significantly at the 5% significance level for both $\hat{\gamma}_{200}$, $F(1, 254) = 3.92$, $p = .05$ and $\hat{\gamma}_{300}$, $F(1, 254) = 3.25$, $p = .007$. The *CP* with values smaller than .93 all have 15 measurement occasions, a large between-study variance ($\sigma^2_{v_3}$= 2.0) and occur when the effect sizes are uncorrected (for both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$). Similar to the constant external event effect, the *CP* is overestimated for $\hat{\gamma}_{300}$ and when the effect sizes are uncorrected in the condition where 30 measurement occasions are included. In the condition where I = 15 and $\sigma^2_{v_3}$= 2.0, the difference between corrected and uncorrected effect sizes is largest.

4.3.2.2 Variance components.

The results are similar to the results of the constant external event effect, and results are less biased using the adjusted model. We only discuss the estimated variances for the immediate treatment effect, because the results are similar for the estimated treatment effect on the trend. When we estimate the between-study variance and the effect sizes are uncorrected, the bias ranges from -0.002 to 3.41, while it ranges from 0.002 to 0.73 for the corrected effect sizes. So the estimated variances depend on the model type [$F(1, 102398) = 1631$, $p < .0001$]. Similar results are obtained for the estimate of the between-case variance. The maximum bias for the corrected effect sizes is 1.60 while it is 3.21 for the uncorrected effect sizes and these estimates depend on the model type, $F(1, 102398) = 5628.62$, $p < .0001$.

4.4 Empirical Illustration

In this section we give empirical illustrations of the comparison of the modified three-level model in which external events are taken into account with the uncorrected model. Therefore, we used a part of the meta-analytic dataset of Heyvaert et al. (2012) in which restraint interventions for challenging behavior among persons with intellectual disabilities was investigated. We give two empirical illustrations of the consequences of ignoring the external event effect in a multiple-baseline across participants design. We illustrate first the consequences of ignoring external events in a single study, and next the consequences of ignoring external events in a three-level meta-analysis.

4.4.1 Ignoring external events in a single study

To illustrate the regression analysis of a multiple-baseline across 3 participants design, we use the study of Thompson et al. (1999) which was included in the meta-analysis of Heyvaert et al. (2012). In their study the effects of benign punishment on the self-injurious behavior of individuals who have been diagnosed with mental retardation was investigated. The three participants were measured repeatedly over time during 22 measurement occasions and the intervention started on session 11, 13 and 20 respectively (see Figure 4.2). From this figure we might expect that there is an immediate reduction in challenging behavior when the treatment is introduced and that the effect of the treatment on the challenging behavior decreases over time (so there is a positive effect on the time trend during the treatment). We also see that the three participants' scores on measurement occasion 4 and 10 are possibly influenced by an external event.

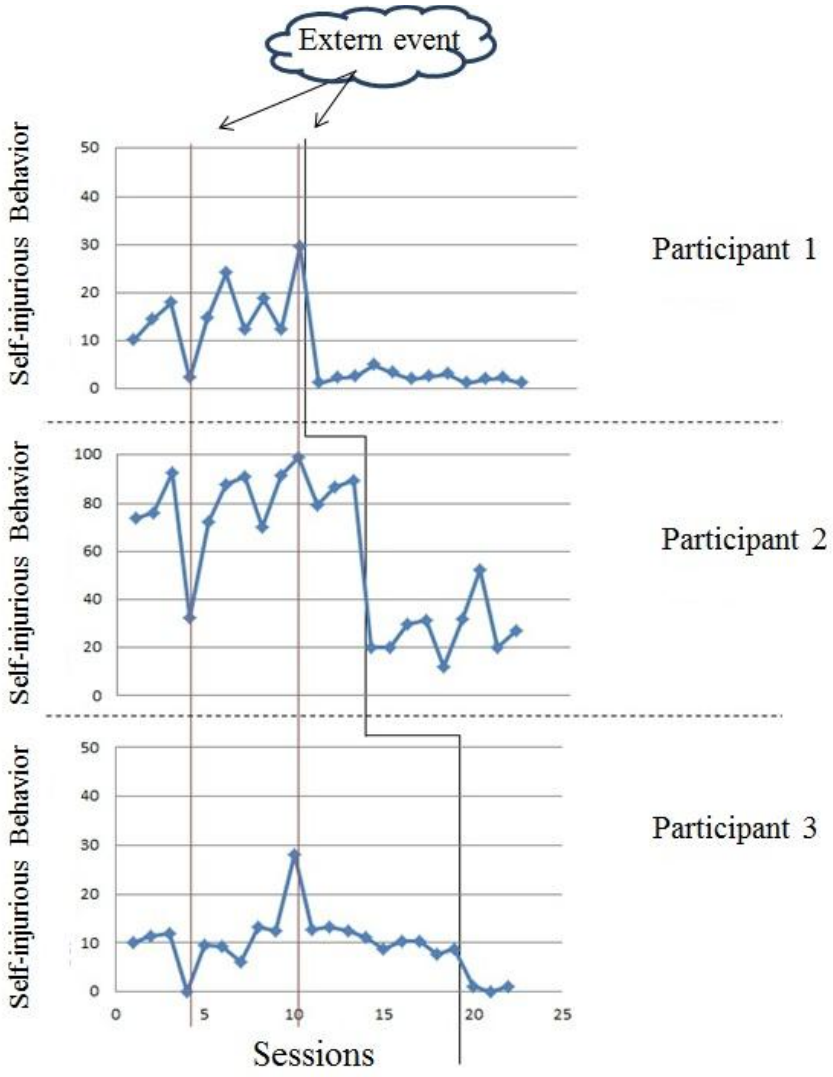

Figure 4.2. Graphical display of a MBD across three participants designs using data from the study of Thompson, Iwata, Conners, and Roscoe (1999).

If we ignore possible external events in the regression analysis before combining the effect sizes in the two-level meta-analyses, the average immediate treatment effect over cases for that study equals -25.58 and the average treatment effect on the time trend over cases from that study equals: -2.58. If we take the external event into account by correcting the effect sizes before combining them, the immediate treatment effect equals -23.23 and the treatment effect on the time trend is 1.24. This means that $\hat{\gamma}_{200}$ is 9.19% smaller when the effect sizes are corrected in comparison with the uncorrected effect sizes. Moreover $\hat{\gamma}_{300}$ is positive for the corrected effect sizes whereas it is negative for the uncorrected which means that the effect of the treatment over time decreases for the corrected effect sizes, whereas it increases for the uncorrected.

4.4.2 Ignoring external events in a three-level meta-analysis

The three-level analysis of SSED data includes summarizing the immediate treatment effect and the treatment effect on the time trend over participants and over studies.

We estimate the immediate treatment effect and the treatment effect on the time trend across seven studies. Again, we use the meta-analysis of Heyvaert et al. (2012) to randomly select multiple-baseline across participants studies. We combined the multiple-baseline across participants study of Lindberg et al. (1999); Chung and Cannella-Malone (2010); Zhou et al. (2000); Thompson et al. (1999); Hanley et al. (2000); Rolider et al. (1991); and Roscoe at al. (1998). In all these studies the same dependent variable was measured, namely the reduction in self-injurious behavior. Again, we compare the three-level meta-analysis of uncorrected and corrected effect sizes.

Results

With the uncorrected effect sizes in the three-level meta-analysis, the average immediate treatment effect equals: -33.14, $t(6.39) = -3.44$, $p = .012$ and the average treatment effect on the time trend equals: -4.42, $t(3.95) = -1.52$, $p = .19$. When correcting the effect sizes before estimating the effects over participants, the immediate treatment effect equals: -21.07, $t(6.88) = -1.13$, $p = .30$, and the treatment effect on the time trend equals: -0.43, $t(1) = -0.28$, $p = .83$. This means that the immediate treatment effect of the corrected effect sizes is 36.42 % smaller compared with the uncorrected effect size and the treatment effect on the time trend for the corrected effect size during the treatment is 90.27% smaller.

This is consistent with the results of the simulation study where we found that the estimated treatment effects are biased when the effect sizes are uncorrected before combining them in the three-level meta-analysis.

4.5 Discussion

4.5.1 General conclusion

External event effects are common in SSEDs because single-case researchers often implement these kinds of designs in everyday scenarios where they cannot control for outside factors (Christ, 2007; Kratochwill et al., 2010; Shadish et al., 2002;). External events are not always anticipated by researchers and thus they may not be measured during the conduct of the study. Furthermore, the size of an event effect may be small and researchers may be unaware of it even after the study has been completed. Whether researchers recognize an external event or not, the failure to account for the event in a meta-analysis can bias the estimate of the treatment effect. Thus, we searched for a method to model external events that could be applied even when the events had not been previously identified. Because we used a multiple-baseline across participants design, there was a need to take into account the interdependence of the participants. Therefore, an external event that influenced the scores of one participant was assumed to influence the scores of the other participants in the same study.

We discussed two possible scenarios. In one scenario, the external event effect remains constant and influences the scores of all participants within a study on four subsequent moments. This occurs for example when a teacher is ill and a substitute teacher takes over the classroom or when a foreign observer is present on subsequent measurement occasions. In the second scenario the external event's effect would likely gradually fade away over four subsequent moments. For instance, the influence of a teacher intern on the behavior of students reduces over time. Moreover, the model adjusted for external event effects takes into account that measurement occasions closer in time are more related than measurement occasions further in time.

We evaluated this approach using a large simulation study and gave some empirical examples. If there is an external event effect of zero, both models (the one that corrects for moment effects and the one that does not) are appropriate. If the external event influences subsequent scores for all the participants within a study, the three-level approach for uncorrected effect sizes is not recommended because the estimates of both treatment effects (i.e., immediate effect on level and effect on time trend) are substantially biased. The *MSE*, standard error, and *CP* are better estimated when using the modified model, which includes moment effects. The difference between the corrected and uncorrected effect sizes is largest when there are a small number of studies and measurement occasions, so in this context we

advise using the adjusted model. Moreover the adjusted model results in less biased variance estimates.

4.5.2 Limitations and suggestions for future research

But of course we should be aware of some limitations. We assumed that all the participants within a study are influenced the same way by the external event effect. It is possible that different participants from the same study are at separate locations and therefore are not all influenced by the external event. Modeling event effects that are not common to all participants in a study is an important avenue for future research.

We chose to keep the number of measurements within a study constant for all participants within the same study. Of course it is possible that different participants of the same study have different series lengths.

Furthermore, we cannot generalize these results to other conditions not involved in this simulation study, but we partially addressed this by simulating a large number of conditions and choosing realistic values for the parameters.

Another limitation is that we assumed linear trajectories in the treatment phase, which might not be true in some real situations. To simplify the simulation model, we further did not account for a possible dependence between regression coefficients, which can be accounted for in a multilevel analysis by estimating the covariance at the various levels.

In addition, subjects in multiple-baseline designs are repeatedly measured, and succeeding measurements may be more related to each other than measurements further away in time. We did not account for this possible autocorrelation and suggest this as a useful extension to the current study.

Kazdin (2010) argued that there needs to be a minimum of three measurement occasions between the participants in a multiple-baseline design in order to show an experimental effect. We did not take this into account in the condition where the number of measurement occasions was 15 because it was not possible to do this and provide each of seven participants a unique baseline. We could alter the intervention schedule to introduce the treatment for some participants (e.g., randomly selected pairs) at the same moment. Examining this strategy specifically, and alternative intervention schedules more generally, would allow further research to extend results to a wider range of multiple-baseline applications.

It can be difficult to attribute simultaneously unusual outcome scores for all participants within a study to an external event effect. If there is no external event effect, we can still use the corrected model because both the corrected and uncorrected effect sizes will be unbiased and thus there is no need to identify before the analyses whether an external event

effect occurred or not. We advise single-case researchers to first use both models in the sensitivity analysis and then decide which model to use. If researchers are interested in the occurrence of external event effects, we recommend that they keep a log in order to identify potential outside factors that may influence the scores at certain measurement occasions and include dummy indicator variables at least for these moments.

The extension of the three-level model for multiple-baseline across participants designs to include modeling of potential external effects makes it even more appropriate and useful for the analysis of realistic SSED datasets. This study has indicated that the three-level model corrected for external event effects provides better results than the uncorrected model for combining results from multiple-baseline across participants data especially if there are only a small number of observations ($I = 15$) and a small number of studies ($K = 10$) in the synthesis. As was found here, even when an external event effect is small, a failure to correct for it can lead to biased effect sizes. Thus, applied SSED researchers are encouraged to consider use of the three-level model that corrects for external event effects when synthesizing results of multiple-baseline design data.

Chapter 5|

The Misspecification of the Covariance Structures in Multilevel Models for Single-Case Data[4]

Abstract

The impact of misspecifying covariance matrices at the second and the third levels of the three-level model on inferences regarding average treatment effects and (co)variances between treatment effects is evaluated by means of a simulation study and an empirical illustration. The results indicate that ignoring an existing covariance has no effect on the treatment effects estimates, but results in underestimation of the variances of the treatment effects and of the standard errors of the treatment effect estimates. If the population covariances are zero, analyses including or not including covariance parameters yield similar results. Single-case researchers are encouraged to use the three-level model including covariances between the treatment effects at the second and third level when synthesizing multiple-baseline design data.

Keywords: Multilevel modeling, multiple-baseline designs, covariance misspecification, Monte Carlo Simulation Study

[4] This chapter has been submitted as a manuscript to the *Journal of Experimental Education* and the revised version is under review: Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S.N.,& Van den Noortgate, W. (2014a). The misspecification of the covariance structures in multilevel models for single-case data: A Monte Carlo simulation study. *Journal of Experimental Education.*

5.1 Introduction

Single-case experimental designs make important contributions to the field of educational research (National Research council, 2002; Odom et al., 2005). For instance, this kind of design can be applied to evaluate specific interventions to reduce challenging behavior in persons with intellectual disabilities or to search for strategies for persons with learning disabilities. Although single-case designs (SCDs) are increasingly popular (Kazdin, 2011), the quantitative analysis of study results obtained with this kind of design is still developing (Kratochwill et al., 2010). The results of a SCD study investigating the effect of an intervention are especially informative for the specific case under investigation, but it is hard to generalize conclusions to other cases. To investigate generalizability of the SCD results across cases, one can collect information for several cases, as is done in the multiple-baseline design (MBD) across cases. In this type of design, an AB phase design is implemented simultaneously to different cases, while the start point of the treatment is staggered (as in Figure 5.1) across cases (Ferron & Scott, 2005; Onghena, 2005; Onghena & Edgington, 2005).

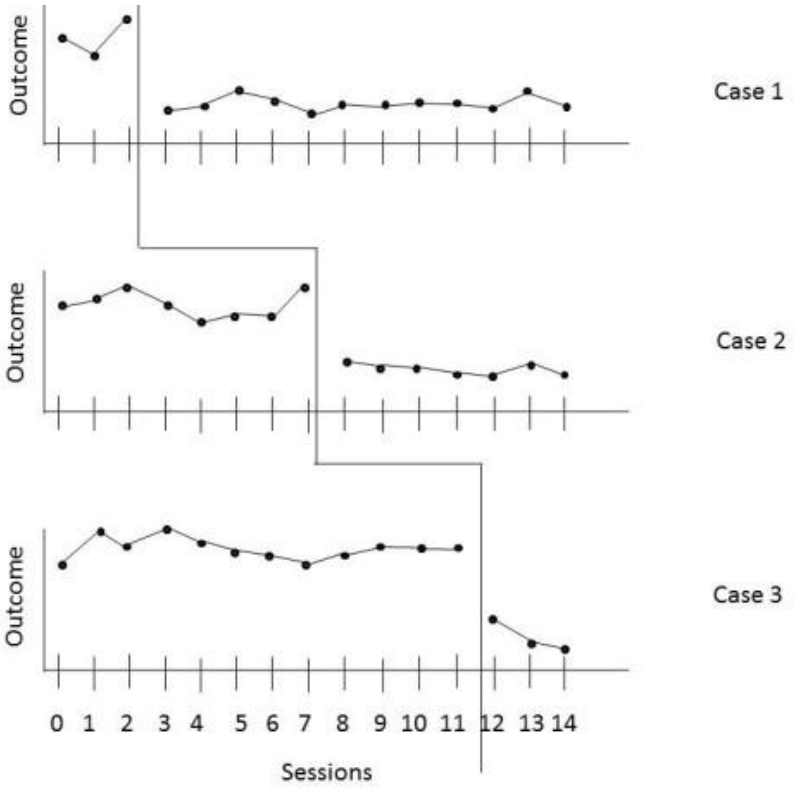

Figure 5.1. Graphical display of the multiple-baseline across participants design using hypothetical data. The start of the intervention is staggered across the three cases.

The MBD is growing in popularity because external events, which are random unexpected events influencing the outcome scores, can be disentangled from treatment effects. These external events might affect the outcome scores of several cases at the same time, while

treatment effects are expected to occur immediately after the treatment starting point which is case-specific (Barlow & Hersen, 1984; Kinugasa et al., 2004; Koehler & Levin, 2000).

5.2 Multilevel Analysis of Multiple-Baseline Across Cases Design

To combine multiple cases' data, multilevel models can be used. Multilevel models are extensions of linear models and make it possible to synthesize treatment effects across cases and studies. When combining SCD data from several MBD studies, a three-level hierarchical structure can be modeled: measurement occasions (i.e., first level units) are nested within cases (i.e., second level units), which in turn are nested within studies (i.e., third level units). For example, consider K studies (k = 0, 1,…, K), with J_k cases in study k (j = 0, 1,…J_k), and I_{jk} measurements for case j from study k (i = 0, 1,…I_{jk}). At level one, the continuous response variable can be modeled, for instance, using an extension of the piecewise regression equation of Center et al. (1985-1986):

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}D_{ijk}T'_{ijk} + e_{ijk} \text{ with } e_{ijk} \sim N(0, \sigma_e^2), \quad (5.1)$$

and the errors, the e_{ijk}'s, are assumed to be independently, identically, and normally distributed. The score on the continuous dependent variable on measurement occasion i for case j from study k (Y_{ijk}) depends on a time-related predictor (T_{ijk}), a binary coded treatment indicator (D_{ijk}), and an interaction term between the time predictor and the dummy variable. The predictor T_{ijk} refers to the measurement occasion of the dependent variable and can be expressed for instance in days or session numbers. For case j from study k, there are n_{Ajk} observations in phase A, n_{Bjk} observations in phase B, so that $n_{Ajk} + n_{Bjk} = I_{jk}$. The second predictor, D_{ijk}, indicates whether the measurement occasion i from case j within study k belongs to the baseline phase (D_{ijk} = 0) or the treatment phase (D_{ijk} = 1). The last term is an interaction term between T_{ijk} and D_{ijk}. In this interaction term, T_{ijk} is centered around its value at the start of the treatment phase, such that the coefficient of the treatment dummy can be interpreted as the immediate effect of the treatment. The centered time variable in the interaction term is indicated by T'_{ijk} and equals $T_{ijk} - (n_{Ajk} + 1)$. This means that T'_{ijk} takes on negative values during the baseline phase and counts down from the first measurement occasion of the treatment phase to the first observation of the baseline phase. The general coding form for case 1 from study 1 is given in Table 5.1.

Table 5.1

Demonstration of General Coding for Case 1 within Study 1

T_{i11}	D_{i11}	T'_{i11} [= $T_{i11} - (n_{ajk}+1)$]	Y_{i11}
1	0	1 - (n_{a11}+1)	Y_{111}
2	0	2 - (n_{a11}+1)	Y_{211}
.	.	.	.
.	.	.	.
.	.	.	.
n_{a11}	0	-1	Y_{n_a11}
n_{a11} + 1	1	0	$Y_{(n_{a11}+1)11}$
n_{a11}+ 2	1	1	$Y_{(n_{a11}+2)11}$
n_{a11} + 3	1	2	$Y_{(n_{a11}+3)11}$
.	.	.	.
.	.	.	.
.	.	.	.
n_{a11}+ n_{b11} = I_{11}	1	$n_b - 1$	$Y_{n_a+n_b11}$

Equation 5.1, regressing Y_{ijk} on T_{ijk}, D_{ijk} and $D_{ijk}T'_{ijk}$ contains four coefficients: β_{0jk} is the intercept and indicates the expected baseline level at the start of the baseline phase (i.e., when all other predictors equal zero), β_{1jk} is the linear trend during the baseline phase, β_{2jk} refers to the immediate treatment effect (i.e., the difference between the estimated outcome score at time zero under the treatment phase and the estimated outcome score at the same point in time under the baseline phase) and β_{3jk} is the effect of the treatment on the time trend for participant *j* in study *k*. Single-case researchers are mainly interested in β_{2jk} and β_{3jk} because they provide information about the change associated with the introduction of the treatment.

At the second level, the variation across cases can be modeled as follows:

$$\begin{cases} \beta_{0jk} = \theta_{00k} + u_{0jk} \\ \beta_{1jk} = \theta_{10k} + u_{1jk} \\ \beta_{2jk} = \theta_{20k} + u_{2jk} \\ \beta_{3jk} = \theta_{30k} + u_{3jk} \end{cases} \text{ with } \begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{2jk} \\ u_{3jk} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0u_1} & \sigma_{u_0u_2} & \sigma_{u_0u_3} \\ \sigma_{u_1u_0} & \sigma^2_{u_1} & \sigma_{u_1u_2} & \sigma_{u_1u_3} \\ \sigma_{u_2u_0} & \sigma_{u_2u_1} & \sigma^2_{u_2} & \sigma_{u_2u_3} \\ \sigma_{u_3u_0} & \sigma_{u_3u_1} & \sigma_{u_3u_2} & \sigma^2_{u_3} \end{bmatrix} \right) \quad (5.2)$$

These equations model that the β coefficients from Equation 5.1 randomly vary across cases, around study-specific means, the θ coefficients. The coefficients along the diagonal of the covariance matrix, $\sigma^2_{u_0}$, $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, and $\sigma^2_{u_3}$, indicate the between-case variance in the intercept, the time trend during the baseline, the immediate treatment effect and the treatment effect on the slope, respectively. The off-diagonal coefficients represent covariances. For instance $\sigma_{u_0u_1}$ indicates the covariance between the intercept and the time trend during the baseline phase.

At the third level, potential variability in the study-specific regression coefficients from the second level equations, the θ coefficients, is modeled. In the fullest model, the θ coefficients each equal an average estimate across studies, indicated by the γ coefficients, and a random deviation from this average:

$$\begin{cases} \theta_{00k} = \gamma_{000} + v_{00k} \\ \theta_{10k} = \gamma_{100} + v_{10k} \\ \theta_{20k} = \gamma_{200} + v_{20k} \\ \theta_{30k} = \gamma_{300} + v_{30k} \end{cases} \text{with} \begin{bmatrix} v_{00k} \\ v_{10k} \\ v_{20k} \\ v_{30k} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{v_0} & \sigma_{v_0 v_1} & \sigma_{v_0 v_2} & \sigma_{v_0 v_3} \\ \sigma_{v_1 v_0} & \sigma^2_{v_1} & \sigma_{v_1 v_2} & \sigma_{v_1 v_3} \\ \sigma_{v_2 v_0} & \sigma_{v_2 v_1} & \sigma^2_{v_2} & \sigma_{v_2 v_3} \\ \sigma_{v_3 v_0} & \sigma_{v_3 v_1} & \sigma_{v_3 v_2} & \sigma^2_{v_3} \end{bmatrix} \right) \quad (5.3)$$

Multilevel modeling entails the advantage that average treatment effects can be estimated, as well as variation between studies and cases in the treatment effect, or study- and case-specific treatment effects. Another major advantage of this multilevel approach is its flexibility. For instance, the model can be extended by including (additional) predictors at each level. Moreover, a specific structure for the variances and covariances at either level can be specified.

Previous research indicates that multilevel modeling works appropriately to combine unstandardized (Moeyaert et al., 2013a) and standardized (Moeyaert et al., 2013b) SCD data across cases and studies. Estimation of the three-level model for SCD data was investigated, by evaluating the estimates of the average immediate treatment effect, γ_{200}, of the average treatment effect on the slope, γ_{300}, and of the between-case and between-study variance of both kinds of effect. However, in these studies, the between-case residuals (u_{0jk}, u_{1jk}, u_{2jk}, u_{3jk}) were each assumed to be independently, identically, and normally distributed with mean zero and homogeneous variance, and thus a diagonal covariance structure was assumed at level-2. However, this might be an over-simplification of the between-case covariance structure. A non-zero covariance between residuals at level 2 seems reasonable, for instance, when due to a ceiling effect the treatment effect is expected to be smaller for cases with an already high baseline level. In addition, these simulation studies made the same assumptions about the between-study residuals (v_{0jk}, v_{1jk}, v_{2jk}, v_{3jk}) but again an unstructured covariance matrix, which allows the level-3 residuals to covary, may be more reasonable than a diagonal covariance structure. To date, no research has focused on the consequences of ignoring truly non-zero covariances in the context of multilevel modeling of SCD data. In most multilevel modeling software, the default option is to estimate an unconstrained covariance matrix for the random effects. However, given that there are four coefficients included in the level-1 regression equation, a total of 21 random effects covariance parameters can be estimated, which complicates estimation especially in scenarios with small sample sizes and possible covariance

values that are close to zero. Therefore there is a need to investigate if estimation of the multilevel model is robust to covariance matrix misspecification. If estimation is reasonably robust, then modeling of a simplified covariance matrix can be recommended for future studies using the three-level model in the context of SCD data.

While only a little research has examined these issues in the context of the multilevel model for SCD data, there is some methodological research that has focused on specification of the residuals' covariance matrix contexts other than SCDs. For example, Singer and Willett (2003) argue that ignoring a covariance in a multilevel model in general may bias the estimation of the standard errors of the average regression coefficients. This will in turn lead to distorted Type I error rates when testing the statistical significance regression coefficients and will affect estimation of the confidence intervals for the effects of interest. Kwok et al. (2007) investigated by means of a simulation study the misspecification of the within-case covariance structure in multiwave longitudinal multilevel models and found that the misspecification has a substantial impact on the variance estimates. Work by Berkhof and Kampen (2004) examines the effect of omitting a random coefficient in the multilevel models in general on the estimated variance components and the estimated variance of the treatment effects. They found that the consequences depend on the between-unit variance proportions. Another study, by Van den Noortgate and Onghena (2005), investigated the effects of ignoring a level from a four level model (in the area of school effectiveness research) on the parameter estimates and standard errors. They found that the variance estimate of the ignored level is divided between the other levels and estimates of the standard errors of the fixed effects and the random components may change.

Specifically for SCD data, the effects of level-1 residuals' covariance misspecification have been studied before for a two-level model (Ferron et al., 2009). In SCDs, it is reasonable that an external variable that influences an observation at a certain moment, also affects succeeding observations. This means that errors from succeeding occasions can be more alike than errors of occasions further in time (Kromrey & Foster-Johnson, 1996). Ferron et al. (2009) found that not modeling autocorrelation in a two-level analysis of SCD data results in too small coverage proportions of the 95% confidence intervals and positively biased variance estimates. This same pattern of results also apply when level-1 residuals' autocorrelation is not modeled for the three-level model (Petit-Bois, Baek, & Ferron, 2012). Level-2 and level-3 covariance misspecification issues in the SCD three-level modeling framework have not yet been investigated. The main focus of this paper is to examine the consequences of level-2 and level-3 covariance matrix

misspecification which should provide a more complete understanding of misspecification issues in contexts of three-level modeling of SCD data.

5.3 Simulation Study

We conducted a simulation study to evaluate estimation of the three-level model when freely estimating covariances between pairs of residuals at levels two and three in the model. It might be possible to mathematically derive large-sample approximations of the estimated standard errors of the treatment effects. However, in the context of multilevel modeling of SCD data, researchers deal with very small sample sizes which violate asymptotic assumption upon which the algebraic derivations would be based. Thus, we exclusively rely on simulation studies to empirically examine estimation of model parameters and standard errors under the realistic sample size values that are typically encountered in applied SCD research in the educational and social sciences.

We simulated raw data using the three-level model in Equations 5.1 through 5.3. To estimate the three-level model parameters, the restricted maximum likelihood procedure in SAS 9.3 PROC MIXED was used (Littell et al., 2006). The Satterthwaite method was used to estimate the degrees of freedom because this method is relatively fast and accurate (Ferron et al., 2009). Convergence was obtained in all conditions and replications.

The criteria used to evaluate the results of the three-level analysis included the bias, the mean squared error (*MSE)*, the standard error (*SE)*, and the coverage proportion (*CP*) of the 95% confidence intervals of the effect parameter estimates. In addition, we looked at the bias of the variance and covariance parameter estimates for both treatment effects.

For this Monte Carlo study, we varied seven design conditions, namely the immediate treatment effect, the treatment effect on the time trend, the number of units at the three levels (i.e., the number of measurements at the first level, the number of cases at the second level and the number of studies at the third level), the between-cases, and the between-studies variability. In order to identify values for the seven design conditions that are authentic for data encountered in the area of educational research, we re-analyzed published meta-analyses of SCD studies (Denis et al., 2011; Heyvaert et al., 2012; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011). Based on the review, the immediate treatment effect on the outcome score, γ_{200}, was given a value of 0 or 2, whereas the treatment effect on the time trend was manipulated to have values of 0 or 0.2. The number of units at the third level (i.e., the number of studies), *K*, was also manipulated because previous simulation studies (Moeyaert et al., 2013a, 2013b; Owens & Ferron, 2012;) indicate that the level-3 units have a significant effect on the results. We

simulated conditions with 10 and 30 studies. While very typical SCD studies tend to have 3, 4 or 7 participants per study, we chose to only investigate 4 and 7 participants per study, because previous simulation studies had found similar results for 3 and 4 cases. Based on our review, we chose to only include one value for the number of measurements within a case, namely 15. The reason for this was twofold. First of all, 15 observation within a case is common (e.g., Shadish & Sullivan, 2011, found that more than 20% of the published SCD studies in 2008 having less than 15 data points) in SCDs and secondly, we want to evaluate the covariance misspecification especially in conditions that are potentially problematic (i.e., a small number of measurement occasions). Based on previous simulation studies (Moeyaert et al., 2013a, 2013b), two other factors potentially influence the estimated treatment effects and variances of these estimated effects, namely the between-study and the between-case variances. Also manipulation of the true value of the covariance between the treatment effects is an important condition. Therefore the covariance matrices, Σ_u and Σ_v were manipulated to have conditions in which there is zero covariance, a moderate covariance, and a large covariance. The covariances were simulated with positive and negative values. The chosen values for the variances are based on those used in the study of Moeyaert et al. (2013a, 2013b).

Level-2 and level-3 errors were generated from a normal distribution using the RANNOR random number generator in SAS. If the between-study variance of the immediate treatment equaled 8 and the between-study variance of the treatment effect on the time trend equaled 0.08, there were five possibilities for the covariance matrix at the third level:

$$\Sigma_v = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 0.08 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0.08 \end{bmatrix} \text{ or } \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 0.08 & 0 & 0 \\ 0 & 0 & 8 & 0.48 \\ 0 & 0 & 0.48 & 0.08 \end{bmatrix} \text{ or } \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 0.08 & 0 & 0 \\ 0 & 0 & 8 & -0.48 \\ 0 & 0 & -0.48 & 0.08 \end{bmatrix} \text{ or } \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 0.08 & 0 & 0 \\ 0 & 0 & 8 & 0.79 \\ 0 & 0 & 0.79 & 0.08 \end{bmatrix}$$

or $\begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 0.08 & 0 & 0 \\ 0 & 0 & 8 & -0.79 \\ 0 & 0 & -0.79 & 0.08 \end{bmatrix}$, representing no covariance, moderate positive and negative covariance and large positive and negative covariance between the immediate treatment effect and treatment effect on the time trend regression coefficients. If the between-study variance of the immediate treatment equaled 2 and the between-study variance of the treatment effect on the time trend equaled 0.2, there were five possibilities for the covariance matrix at the third level:

$$\Sigma_v = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix} \text{ or } \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 2 & 0.38 \\ 0 & 0 & 0.38 & 0.2 \end{bmatrix} \text{ or } \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 2 & -0.38 \\ 0 & 0 & -0.38 & 0.2 \end{bmatrix} \text{ or } \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 2 & 0.63 \\ 0 & 0 & 0.63 & 0.2 \end{bmatrix} \text{ or}$$

$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 2 & -0.63 \\ 0 & 0 & -0.63 & 0.2 \end{bmatrix}$, representing no covariance, moderate positive or negative covariance and

large positive or negative covariance between the immediate treatment effect and treatment effect on the time trend regression coefficients. The same values were chosen for the between-case covariance matrix. The within-case variance was generated with a variance of one and assumed to be homogeneous across phases.

For simplicity, we matched the covariance generating values' pattern used at level 2 with that used at level 3. For example, for conditions where a moderate covariance was used to generate level-2 residuals, the level-3 residuals were also generated using a moderate value. The same held for zero covariance conditions and for large covariance conditions. In addition, we matched the direction of the covariance between residuals for level-2 and level-3 covariances.

As is common in MBD, we staggered the introduction of the intervention across cases within studies. The staggering is a function of the total number of cases (see Table 5.2).

Table 5.2

Staggering of the Intervention's Start Point as a Function of the Number of Cases (J)

J	*J*	Start of treatment
4	1	3
	2	6
	3	9
	4	12
7	1	3
	2	6
	3	6
	4	9
	5	9
	6	12
	7	12

We examined a total of $2^6 = 64$ conditions and for each condition we simulated 500 datasets. Because the simulation was computationally very intensive, we choose to simulate 500 replications and included a large number of conditions instead of simulating a larger number of replications and including a small number of conditions. There were 32,000 datasets to analyze. After each dataset was generated, the simulated dataset was analyzed using a three-level multilevel model with maximum likelihood estimation via the MIXED procedure in SAS. Each dataset was analyzed twice, once, assuming a model that freely estimated covariance components, and another time assuming a model that constrained covariances to zero. In the model that included the covariances, the variances of the four random effects at both of the higher levels were estimated together with the covariances between the treatment effects. In addition, the within-case variability was estimated. This means that there were a total of 11

random effects variance component parameters that were estimated. When the covariance components were not included, nine variance component parameters were estimated.

This results in four different scenarios: A1, A2, B1, and B2, see Table 5.3. The letter (i.e., A and B) indicates the model used to generate the data. In situation A, a zero covariance was generated whereas in situations B non-zero covariance parameters were used to generate the data. The number (i.e., 1 and 2) refers to the analysis model. The models accompanied with number 1 indicate that the data were analyzed assuming zero covariance whereas number 2 refers to an analysis model in which covariance parameters were estimated. We expect that situations A1, A2 and B2 will give us approximately correct estimates of the average treatment effects and the (co)variances in treatment effects, but are especially interested in problems due to model misspecification (situation B1).

Table 5.3

Four Combinations of Generating and Estimating Models

		Model Used to Generate Data	
		Without covariance	With covariance
Model to analyze the data	Without covariance	Scenario A1	**Scenario B1**
	With Covariance	Scenario A2	Scenario B2

5.4 Results

We are especially interested in potential differences in results between scenarios A1 and A2 and between scenarios B1 and B2. We only discuss the results for $\hat{\gamma}_{200}$ because similar results were obtained for $\hat{\gamma}_{300}$. Moreover, we only report the results for conditions in which a large positive covariance was generated in scenarios B1 and B2, because the effects of ignoring a moderate covariance are similar. Also similar results are found independent of the sign of the covariance between the treatment effects.

5.4.1 Average immediate treatment effect

5.4.1.1 Bias and mean squared error

We first look at the bias, which is generally defined as the mean difference between the estimated values and the population value. Figure 5.2 presents the deviations of the estimated immediate treatment effect from the population value. We can conclude that the mean deviation

is small, but that the variance of the deviations is larger when there is covariance in the generated data (i.e. scenario B1 and B2). The mean bias over all conditions is small and independent of the scenario. It equals –0.0070, –0.0022, –0.0015 and 0.0012 for scenario A1, B1, A2 and B2 respectively. The number of studies (*K*), the number of cases (*J*), the true values of the between- and within study variances ($\sigma^2_{v_2}$ and $\sigma^2_{u_2}$) have no significant influence on the estimated bias.

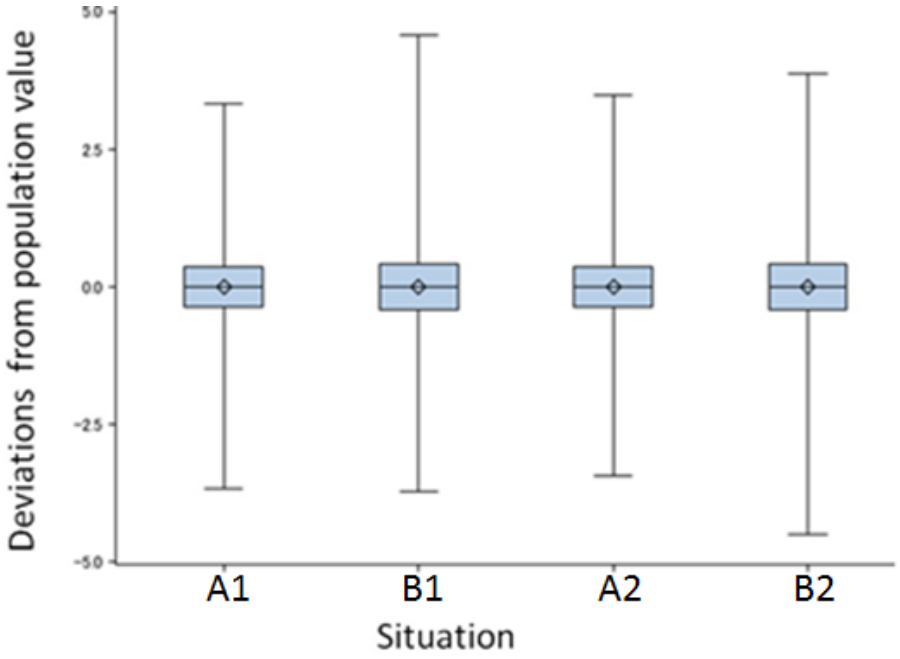

Figure 5.2. Distribution of the deviations of the estimated immediate treatment effect from its population value (γ_{200}). The length of the box represents the interquartile range. The symbol and the horizontal line in the box interior indicate the scenario mean and median, respectively. The whiskers issuing from the box extend to the scenario minimum and maximum values.

Table 5.4 gives an overview of the estimated *MSE* by condition and scenario. Based on Figure 5.2, we expect no difference in *MSE* between the two scenarios in which zero covariance is generated and the two scenarios in which non-zero covariance is generated which is supported by the ANOVA results (between scenarios A1 and A2, $F(1,126) = 0.00$, $p = 1.00$ and between scenario B1 and B2, $F(1,126) = 0.02$, $p = .88$. This means that if the *MSE* of estimates of the immediate treatment effect is taken as a criterion, it does not matter whether the analysis model includes covariance parameters. As expected from Figure 5.2, the *MSE* is larger in scenarios B1 and B2 compared to scenarios A1 and A2, $F(1,126) = 2.89$, $p = 0.09$. Across all conditions, the ANOVA also indicates that for all scenarios, a large number of studies and cases and a small within- and between-study variances result in a smaller *MSE*.

Table 5.4

The MSE of $\hat{\gamma}_{200}$; for γ_{200} = 2, γ_{300} Conditions for the Four Scenarios

			J = 4		J = 7	
Scenario	K	$\sigma_{v_2}^2$	$\sigma_{u_2}^2$ = 2	$\sigma_{u_2}^2$ = 8	$\sigma_{u_2}^2$ = 2	$\sigma_{u_2}^2$ = 8
A1	10	2	0.16	0.20	0.11	0.15
		8	0.38	0.41	0.34	0.38
	30	2	0.05	0.07	0.04	0.05
		8	0.11	0.17	0.12	0.14
A2	10	2	0.15	0.22	0.11	0.15
		8	0.35	0.44	0.39	0.45
	30	2	0.05	0.06	0.04	0.05
		8	0.14	0.18	0.12	0.14
B1	10	2	0.15	0.26	0.15	0.19
		8	0.58	0.62	0.45	0.54
	30	2	0.06	0.08	0.04	0.07
		8	0.16	0.22	0.17	0.15
B2	10	2	0.17	0.30	0.13	0.16
		8	0.50	0.74	0.41	0.51
	30	2	0.08	0.08	0.04	0.07
		8	0.19	0.22	0.13	0.15

5.4.1.2 Estimates of standard error and coverage proportion of the 95% confidence interval

The standard errors of the treatment effect estimates are used to construct confidence intervals around the estimated treatment effects, $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. The standard deviations of the effect estimates in a given condition can be used as an empirical approximation of the true standard error and therefore as a criterion to evaluate the standard error estimates. We look at the relative standard error bias, which is the difference between the median standard error estimate and the standard deviation of the estimate of the effect divided by the standard deviation of the estimate of $\hat{\gamma}_{200}$ (Hoogland & Boomsma, 1998).

The values for the relative standard error biases are negative (see Table 5.5), which means that the standard error estimates are smaller than expected. There is no significant difference in the relative standard error bias between scenarios A1 and A2, $F(1, 126) = 0.02$, $p = .89$ and between scenarios B1 and B2, $F(1, 126) = 3.46$, $p = .07$, but the median relative standard error biases in scenarios B1 (= -0.17) and B2 (= -0.14) are considered substantial (more than 10%, Hoogland & Boomsma, 1998) in comparison with scenarios A1 (-0.024) and A2 (-0.023): $F(1,126) = 362.21$, $p < .0001$. The ANOVA also indicated that *K* has a substantial effect on the standard error in scenarios A1 and A2, whereas *K*, *J*, $\sigma_{u_2}^2$ and $\sigma_{v_2}^2$ have an effect on the standard error in scenarios B1 and B2.

The difference in *CP* of the 95% confidence intervals between scenarios A1 and A2 is not significant, $F(1, 126) = 0.03$, $p = .86$, whereas the difference between scenarios B1 and B2 is significant, $F(1, 126) = 10.51$, $p = .002$. In scenarios A1 and A2, the mean *CP* equals .95 as expected. The mean *CP* in scenarios B1 and B2 equal .91 and .92, respectively. The conditions with too-small *CP*s correspond to the conditions with too-small standard error estimates (see Table 5.5). In general, the results are slightly better in scenario B2 in comparison to scenario B1. In scenario B1 and B2, respectively 84 % and 75% of the conditions have a *CP* smaller than .93, whereas this is 0% and 3% in scenarios A1 and A2. The *CP* in scenarios B1 and B2 is not influenced by the parameters, whereas the *CP* in the other scenarios becomes closer to the nominal level of .95 if more cases ($J = 7$) and a smaller between-case and between-study variance is included.

Table 5.5

The Relative Standard Error Biases and the Coverage Proportion of the 95% Confidence Interval for $\hat{\gamma}_{200}$; for γ_{200} = 2 and γ_{300} = 0.2 Conditions

			Relative standard error biases				Coverage proportion			
			J = 4		*J* = 7		*J* = 4		*J* = 7	
Scenario	*K*		$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8
A1	10	$\sigma^2_{u_2}$ = 2	-0.05	-0.03	-0.06	-0.02	0.95	0.94	0.95	0.95
		$\sigma^2_{u_2}$ = 8	-0.04	-0.01	0.02	0.00	0.95	0.94	0.95	0.96
	30	$\sigma^2_{u_2}$ = 2	-0.03	0.06	-0.07	-0.02	0.94	0.96	0.93	0.95
		$\sigma^2_{u_2}$ = 8	-0.04	-0.03	-0.06	0.02	0.93	0.94	0.94	0.95
A2	10	$\sigma^2_{u_2}$ = 2	-0.04	-0.04	-0.01	-0.08	0.95	0.97	0.95	0.94
		$\sigma^2_{u_2}$ = 8	-0.06	-0.03	-0.06	-0.08	0.95	0.93	0.94	0.93
	30	$\sigma^2_{u_2}$ = 2	-0.05	-0.03	0.01	-0.01	0.93	0.95	0.94	0.94
		$\sigma^2_{u_2}$ = 8	-0.05	-0.01	-0.05	-0.04	0.93	0.95	0.94	0.94
B1	10	$\sigma^2_{u_2}$ = 2	**-0.17**	**-0.18**	**-0.13**	**-0.15**	**0.91**	**0.91**	0.94	**0.92**
		$\sigma^2_{u_2}$ = 8	**-0.22**	**-0.19**	**-0.18**	**-0.19**	**0.91**	**0.91**	**0.91**	**0.90**
	30	$\sigma^2_{u_2}$ = 2	**-0.20**	**-0.12**	**-0.15**	**-0.10**	**0.89**	**0.92**	**0.92**	0.94
		$\sigma^2_{u_2}$ = 8	**-0.19**	**-0.23**	**-0.14**	-0.09	**0.90**	**0.87**	**0.92**	0.93
B2	10	$\sigma^2_{u_2}$ = 2	**-0.22**	-0.09	**-0.10**	**-0.12**	**0.92**	0.96	0.94	0.93
		$\sigma^2_{u_2}$ = 8	**-0.22**	**-0.23**	**-0.13**	**-0.13**	**0.91**	**0.92**	**0.91**	**0.91**
	30	$\sigma^2_{u_2}$ = 2	**-0.22**	**-0.14**	-0.08	-0.06	**0.89**	**0.91**	0.93	0.93
		$\sigma^2_{u_2}$ = 8	**-0.24**	**-0.11**	**-0.15**	**-0.09**	**0.88**	**0.93**	**0.92**	0.93

Note. For the relative standard error biases, values smaller than or equal to 0.10 in magnitude appear in bold. For the *CP*, values smaller than .93 and larger than .97 appear in bold.

5.4.2 Variance components estimates

The between-study and between-case variances were estimated for both the immediate treatment effect and the treatment effect on the trend. The results are very similar for a moderate and large covariance. Also similar results are found independent of the sign of the covariance between the treatment effects. Therefore, we only report the results of a large positive generated covariance in scenario B1 and B2. Because variance estimates are positively skewed (skewness = 1.74), due to truncation of negative estimates to zero, we calculated the median (relative) deviation of the estimates from the population value, rather than the mean (relative) deviation, to evaluate the (relative) bias in the estimates.

5.4.2.1 Bias of between-case and between-study variance estimate for the immediate treatment effect

Table 5.6 shows that there is negative relative bias in the estimated between-case variance ($\sigma^2_{u_2}$) in all the conditions and all the scenarios. This means that the estimates are smaller than expected. The relative bias equals -1%, 1%, -27, and -1% in scenario A1, A2, B1, and B2 respectively. This indicates that there is a significant larger amount of relative bias (-27%) in scenario B1, where the covariance matrix is misspecified (see Table 5.6). The bias ranges from -13% to -50% in scenario B1, while the range is much smaller in the other scenarios (from 0% to -8%). The ANOVA indicates that no parameter has an influence on the estimated $\sigma^2_{u_2}$ in scenarios A1, A2 and B2, but that the estimated $\sigma^2_{u_2}$ in scenario B1 can be reduced by decreasing the between-case variance. Similar results are obtained for the estimated between-case variance of the treatment effect on the time trend.

Also the median of relative deviations for the estimated between-study variances for the immediate treatment effect are negative, indicating that the estimates are smaller than expected. The median relative bias is also larger in scenario B1 (-9%) in comparison to the other scenarios (-5%). Moreover the difference in median relative bias between scenario B1 and B2 is obvious, $F(1, 126) = 10812.5$, $p < .001$. In scenario B1 and B2, no parameter seems to influence the estimated $\sigma^2_{v_2}$. However, in scenario A1, the $\hat{\sigma}^2_{v_2}$ can be further reduced by including a large number of cases ($J = 7$) and a small amount of $\sigma^2_{u_2}$ ($\sigma^2_{u_2} = 2$). Furthermore, the $\hat{\sigma}^2_{v_2}$ in scenario A2 is smaller if a large amount of studies ($K = 30$) and cases ($J = 7$) and a small number of between-case and between-study variance is included. Also similar results are obtained for the estimated between-study variance of the treatment effect on the time trend.

Table 5.6

Median of Relative Deviation of the Variance Estimates of γ_{200}, for $\gamma_{200} = 2$ *and* $\gamma_{300} = 0.2$ *Conditions*

			$\hat{\sigma}^2_{u_2}$				$\hat{\sigma}^2_{v_2}$			
			K = 10		*K* = 30		*K =10*		*K = 30*	
Scenario	*J*	$\sigma^2_{u_2}$	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8	$\sigma^2_{v_2}$ = 2	$\sigma^2_{v_2}$ = 8
A1	4	2	-0.05	-0.08	-0.03	-0.02	0.01	-0.07	-0.05	-0.02
		8	-0.04	-0.02	0.00	0.00	-0.11	-0.11	-0.03	-0.01
	7	2	-0.03	-0.05	-0.01	0.00	-0.10	-0.07	-0.03	-0.06
		8	0.01	0.00	-0.01	0.00	-0.08	-0.10	-0.04	-0.03
A2	4	2	-0.06	-0.08	0.01	-0.04	-0.07	**-0.14**	-0.03	-0.02
		8	-0.04	-0.03	-0.02	-0.01	**-0.17**	-0.01	-0.01	-0.01
	7	2	-0.01	0.02	-0.02	-0.01	-0.04	-0.09	-0.02	-0.03
		8	-0.01	-0.01	0.00	-0.01	**-0.11**	**-0.12**	-0.06	-0.02
B1	4	2	**-0.47**	**-0.50**	**-0.44**	**-0.43**	**-0.12**	**-0.15**	-0.06	-0.07
		8	**-0.18**	**-0.18**	**-0.16**	**-0.13**	**-0.20**	**-0.18**	-0.02	-0.09
	7	2	**-0.46**	**-0.45**	**-0.45**	**-0.43**	-0.09	**-0.15**	-0.09	-0.06
		8	**-0.18**	**-0.15**	**-0.15**	**-0.15**	**-0.11**	**-0.15**	-0.07	-0.06
B2	4	2	-0.06	-0.03	-0.02	-0.02	**-0.18**	-0.06	-0.01	-0.03
		8	-0.06	-0.03	-0.01	0.00	**-0.14**	**-0.11**	-0.04	-0.01
	7	2	-0.03	-0.02	-0.01	0.00	**-0.12**	**-0.13**	-0.04	-0.04
		8	-0.01	-0.01	-0.01	-0.01	-0.07	**-0.11**	-0.06	-0.04

Note. Relative median deviations of the variance estimates smaller than or equal to –0.10 appear in bold.

The Scheffé's test for differences between the four scenarios in relative bias indicates that the relative bias in scenario B1 is significantly larger than the biases in the other scenarios (see Table 5.7) and this for both $\hat{\sigma}^2_{u_2}$ and $\hat{\sigma}^2_{v_2}$.

Table 5.7

Scheffé's Test for Differences between Scenarios in relative Biases for $\hat{\sigma}^2_{u_2}$ and $\hat{\sigma}^2_{v_2}$

Comparison Scenarios	$\hat{\sigma}^2_{u_2}$			$\hat{\sigma}^2_{v_2}$		
	Difference	Simultaneous 95% confidence limits		Difference	Simultaneous 95% confidence limits	
		Lower	Upper		Lower	Upper
A1 - B1	**0.015**	0.004	0.265	**-0.039**	0.028	0.051
A1 - A2	-0.009	-0.021	0.003	-0.007	-0.019	0.005
A1 - B2	0.002	-0.010	0.135	0.001	-0.011	0.013
B1 - A2	**-0.023**	-0.035	-0.012	**-0.047**	-0.058	-0.035
B1 - B2	**-0.013**	-0.245	-0.002	**-0.039**	-0.051	-0.027
A2 - B2	0.011	-0.000	0.023	0.007	-0.041	0.020

Note. A significant differences between the scenarios (α = .05) appear in bold

5.4.2.2 Covariance components estimates

Bias of covariance between u_2 and u_3 ($\sigma_{u_2u_3}$) and v_2 and v_3 ($\sigma_{v_2v_3}$) estimates.

The covariance between the residuals at level 2 ($\sigma_{u_2u_3}$) and level 3($\sigma_{v_2v_3}$) were estimated in scenario A2 en scenario B2. The covariances at level 2 and level 3 in scenario A2 are expected to equal zero in all conditions. The mean bias for $\hat{\sigma}_{u_2u_3}$ equals $3.0*10^{-6}$ and for $\hat{\sigma}_{v_2v_3}$, it equals $7.3*10^{-4}$. Also in scenario B2, the covariance estimates are close to their expected value (zero, (-)0.79, (-)0.48, (-)0.63 or (-)0.38). The mean bias for $\hat{\sigma}_{u_2u_3}$ equals -0.001 and the mean bias for $\hat{\sigma}_{v_2v_3}$ equals -0.0005.

5.5 Empirical Illustration

We use the meta-analysis of single-case studies conducted by Denis et al. (2011). They collected studies where the effectiveness of a treatment for self-injurious behavior in people with profound intellectual disabilities was investigated. In particular, 18 studies were collected where non-aversive, non-intrusive forms of reinforcement were examined. We analyzed the data by modeling possible covariance between the regression coefficients (scenario B2) and ignoring possible covariance between the regression coefficients (scenario B1). The SAS codes used to estimate the immediate treatment effect (i.e., γ_{200}), the treatment effect on the time trend (i.e., γ_{300}), the variance components ($\sigma^2_{u_2}$, $\sigma^2_{u_3}$, $\sigma^2_{v_2}$, $\sigma^2_{v_3}$) and covariance components ($\sigma^2_{u_{2,3}}$and $\sigma^2_{v_{2,3}}$) are presented in Addendum A2.

The results indicate that, as expected, the ignorance of existent covariance has no large effect on the estimated treatment effects. The immediate treatment effect estimates are statistically significant at the .01 level and equal -2.66, $t(18.4) = -4.01$, $p = .0008$, and -3.00, $t(18) = -4.23$, $p = .0005$, for scenario B1 and scenario B2 respectively and the estimated treatment effects on the time trend are not statistically significant and equal -0.045, $t(12.5) = -1.01$, $p = .33$, and -0.100, $t(17) = -1.92$, $p = .07$, for scenario B1 and B2 respectively. However, there is a difference in terms of the estimated variance components. The estimated between-study variances of the immediate treatment effect are smaller in scenario B1: $\hat{\sigma}^2_{v_2} = 5.89$, $Z = 2.03$, $p = .02$, in comparison to scenario B2: $\hat{\sigma}^2_{v_2} = 7.65$, $Z = 2.13$, $p = .02$, and none of them are statistically significant at the .01 significance level. Similar results are obtained for the estimated between-study variance on the time trend: $\hat{\sigma}^2_{v_3} = 0.017$, $Z = 1.52$, $p = .06$, in scenario B1 and $\hat{\sigma}^2_{v_3} = 0.018$, $Z = 1.53$, $p = .06$, in scenario B2. This could indicate that the between-study variance is underestimated if we ignore covariance at level 3, a result that is in line with the simulation results. We also identified a difference between the estimated between-case variance of the immediate treatment effect, which equals 2.38, $Z = 1.92$, $p = .02$, in scenario B1 and 3.51, $Z = 2.39$, $p = .008$, in scenario B2. Also a difference in between-case variance of the treatment effect on the time trend was found: $\hat{\sigma}^2_{u_3} = 0.00029$, $Z = 0.18$, $p = .43$, in scenario B1 and $\hat{\sigma}^2_{u_3} = 0.044$, $Z = 2.05$, $p = .02$, in scenario B2. The estimated covariance, in scenario B2, between the immediate treatment effect and the treatment effect on the time trend equals 0.34, $Z = 2.25$, $p = .02$, and -.13, $Z = -0.79$, $p = .43$, respectively at level 2 and level 3. This means that a large immediate treatment effect at level 2 goes together with a large treatment effect on the time trend. At level 3, a large immediate treatment effect means a smaller treatment effect on the time trend.

This empirical example confirms that the estimated variance components of the between-case and between-study variance of the immediate treatment effects depend on the estimated model. The model misspecification could indicate that the between-study variance is underestimated.

5.6 Discussion

5.6.1 General conclusion

The main purpose of this study was to evaluate the consequences of misspecifying the between-case and between-study covariance matrix on the estimation of the treatment effects, and their corresponding mean squared error, standard errors, coverage proportion of the 95% confidence interval and variance and covariance components. Because it is not always obvious how to define the covariance matrix, it is important to examine the degree to which the treatment effect estimates and the variance estimates are sensitive to changes in the specification of the covariance matrix. Therefore, we compared the condition where covariance is simulated, but ignored in the analysis with the scenario where covariance is simulated and estimated in the analysis, and compare the scenario where covariance is not simulated and not estimated with the scenario where covariance is not simulated, but is estimated in the analysis.

As expected from previous research, the results indicate that the average treatment effects estimates are unbiased. The *MSE* is largest in scenario B1 and B2, but is smaller if the number of studies and cases are large and if the between-study and between-case variance is small. The median relative standard error biases difference in scenario B1 and B2 are substantial and only slightly larger in scenario B1, which confirms previous research about multilevel models in general. This in turn results in a too small coverage proportion of the 95% confidence intervals. This indicates that the treatment effect estimates are relatively robust for ignoring covariance. As expected, causes the misspecification in the random part of the multilevel model biased variance estimates. In scenario B1, the estimated between-study variance and between-case variance has extremely large relative bias values for both estimated treatment effects (going up to 27%). In the other scenarios, the variance estimates are unbiased. If there is no covariance in the data, the results are similar for the analyses including covariance or ignoring variance. Thus, this study motivates to model covariance in the analysis model.

5.6.2 Limitations and suggestions for future research

As with any simulation study, one of the major potential limitations of this study is the generalizability of the findings. Further research is needed for the applicability of current findings to a broader range of conditions. We partly addressed this limitation by including realistic conditions based on several re-analysis of meta-analysis. The conditions are quite representative for the research field of single-case experiments in educational settings.

In current research, we only investigated the basic multiple-baseline design including two predictors at the first level. We excluded models with multiple predictors at level 2 and level 3, models using unbalanced data, non-linear models, reversal and alternating designs, and other complex models.

In this study, we only included covariance at the second and third level, which means that we ignored possible autocorrelation at the first level. The issue of autocorrelation itself deserves separate research and is beyond the scope of this paper (Baek & Ferron, 2013). Moreover, we only generated covariance between the regression coefficients indicating treatment effects because the other regression coefficients were set on zero in order to make the estimated treatment effect estimates better interpretable. We nevertheless believe that exploring relative simple scenarios is a first step for a thoughtful study of more complex scenarios, and for a correct interpretation of the results for these more complex scenarios.

The combination of SCD data over studies may be difficult if studies are too different. Studies may for instance differ in measuring the treatment effect. We can handle this by the inclusion of covariates indicating certain study and even case characteristics to model this heterogeneity. Another possibility is standardizing the data or using a multivariate three-level model.

Other approaches to estimate the treatment effects and variances in these treatment effects when the variance structures are misspecified should be considered in future research, such as the sandwich estimator (i.e., cluster-robust or Huber estimators). Even when the covariance matrices are misspecified, the sandwich estimator is asymptotically consistent (Hedges, Tipton, & Johnson, 2010; Raudenbush & Bryck, 2002;). It would be a useful contribution to compare the standard errors and coverage proportion of the 95% confidence intervals constructed with the sandwich estimator to those constructed using the model-based estimators in the misspecified model.

Furthermore, the misspecification of the covariance matrix is only one aspect to test the robustness of the three-level modeling approach. Further research is needed to evaluate other issues such as non-normal data and not identical distributed errors. Meanwhile, we advise single-case researchers to consider use of the three-level model that takes into account covariance when synthesizing results of multiple-baseline design data. If there is no covariance and we use this model, there are no problems, but if we ignore existent covariance, variance components estimates can be seriously biased.

PART 2| APPLICATIONS

Chapter 6|
The Influence of the Design Matrix on Treatment Effect Estimates in the Quantitative Analyses of Single-Case Experimental Design Research[5]

Abstract

The quantitative methods for analyzing single-subject experimental data have expanded during the last decade, including the use of regression models to statistically analyze the data, but still a lot of questions remain. One question is how to specify predictors in a regression model in order to account for the specifics of the design and estimate the effect size of interest. These quantitative effect sizes are used in retrospective analyses and allow synthesis of single-subject experimental study results which is informative for research and policy. We discuss different design matrices that can be used for the most common single-subject experimental designs, namely, the multiple-baseline designs, reversal designs, and alternating treatment designs and provide empirical illustrations. The purpose of this article is to guide single-subject experimental data analysts interested in analyzing and meta-analyzing single-subject experimental design data.

Keywords: single-subject experimental design, piecewise regression equation, multiple-baseline design, reversal design, alternating treatment design, design matrix

[5] This chapter has been accepted for publication in *Behavior Modification*: Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S.N.,& Van den Noortgate, W. (2014b). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental designs research. *Behavior Modification*.

6.1 General Introduction

A single-subject experimental design (SSED) is identified by three important features: (1) data are gathered, analyzed and interpreted for one case (this case can be one participant or a group of participants e.g., a classroom), (2) the participant(s) is (are) observed repeatedly during baseline(s) and treatment(s) phase(s), and (3) outcomes during and after the treatment are compared with outcomes prior to treatment (Kratochwill et al., 2010). The main focus of this design lies in assessing whether there is a causal relation between the introduction of a treatment and the change in a dependent variable (Levin et al., 2003). The demonstration of experimental control is also very important (i.e., a change in outcome scores is due to the introduction of the treatment and not to some extraneous variables). An SSED researcher wants to investigate whether a specific treatment works for a specific subject or group of subjects. Because the subjects in SSEDs are observed repeatedly over time, the time variable plays an important role.

Shadish and Sullivan (2011) investigated 809 SSEDs reported in 2008 in the field of psychology and education and characterized all SSED variants. They used the typology presented in the What Works Clearinghouse Standards (*WWCs*) for SSEDs (Kratochwill et al. 2010) to code the types of designs used. The most popular designs were: multiple-baseline designs (used in 54.3% of the 809 retrieved studies), reversal design (8.2%), and alternating treatment designs (8%). These designs involve phase repetition and therefore handle major threats to internal validity including, for instance, history or maturation (Shadish et al., 2002). These SSEDs have become increasingly popular and are applied in a wide array of research fields such as education, clinical psychology, school psychology, special education, etc. (Franklin, Allison, & Gorman, 1997; Ittenbach & Lawhead, 1997; Wacker, Steege, & Berg, 1988), but the methodology to statistically analyze these kinds of data remains limited.

In the area of SSED research there has been a long tradition of visual inspection of the data during data collection. Visual analysis techniques have long been acknowledged as effective and valuable (Michael, 1974). During visual analysis of the data, the effect of the independent variable and extraneous variables are evaluated while the SSED is being conducted. This ongoing process of data evaluation allows the applied SSED researcher to be responsive to the needs of the subject under investigation (Barlow & Hersen, 1984; Kazdin, 2011). For instance, the intervention can be adapted during observation or the intervention can be introduced only after a stable baseline pattern emerged. This is also known as response-

guided experimentation. Another advantage of visual inspection is that the influence of extraneous variables can be eliminated because of experimental control. Kahn et al. (2010) suggest that visual inspection can lead to consistent interpretation of SSED data among well-trained raters, and Ferron and Jones (2006) demonstrate how Type I error control can be ensured in visual analyses. However, visual analyses by themselves are not well suited for synthesizing literature, limiting the capacity to objectively evaluate an evidence base (DeProspero & Cohen, 1979; Ottenbacher, 1992). The summary of SSED research findings is timely as the number of published SSED studies is increasing at an exponential rate in behavior research areas (e.g., Social Science Citation Index). The regression-based approach discussed in this article is a flexible technique used to analyze SSED data retrospectively, and as a complement to visual analysis during data collection. The purpose of the regression analysis is to quantify the SSED data results using an effect size estimate which can be used to compare SSED results across studies, enhances the communication among applied SSED researchers, and can be used in meta-analysis to synthesize a large body of research. The *WWC* standards (Kratochwill et al. 2010) also recommend combining results from SSED studies because they can provide a strong basis for causal inferences (Horner et al., 2005). Confidence in the validity of treatment effects demonstrated within subjects is enhanced by replication of effects across different subjects, studies, and research groups (Horner & Spaulding, in press). The evidence-based movement in SSED context has emphasized the need for quantitative summaries of the results, especially for making them available for meta-analytic purposes (Jenson, Clark, Kircher, & Kristjansson, 2007).This quantification is needed to contribute to evidence based research and to inform research and practice. By using statistical analysis techniques, the following research questions can be resolved: (1)What is the average treatment effect estimate across studies; (2) what is the magnitude of variation between subjects in the size of the effect?, and (3) What is the influence of a predictor on the treatment effect? To conduct a meta-analysis of SSED studies, researchers have relied on effect sizes. For an in-depth discussion, we refer to Maggin et al. (2011). Nonparametric effect sizes for SSED research, such as the family of non-overlap metrics (e.g., percentage of non-overlapping data, Scruggs, Mastropieri, & Casto, 1987, and percentage of all non-overlapping data, Parker, Hagan-Burke, & Vannest, 2007) do not require distributional assumptions but have been criticized for their inability to (1) account for data trends, (2) discriminate between large treatment effects due to ceiling effects, and (3) lack of a known sampling distribution (Wolery et al., 2010). Use of parametric effect sizes addresses these critiques. In addition, unlike nonparametric effect sizes, parametric effect sizes allow

researchers to calculate interval estimates of treatment effects and to estimate the variability in treatment effects within and across subjects. Regression-based analyses can also be used to examine explanation of variance by predictors such as age, gender, SES, school type, etc. (Van den Noortgate & Onghena, 2003b).

In this paper, we focus on one specific parametric effect size obtained through regression analysis of SSED data (Allison & Gorman, 1993; Beretvas & Chung, 2008; Center et al., 1985-1986; Huitema & McKean, 2000; Kratochwill & Levin, 2010; Maggin et al., 2011). Davis, Gagné, Frederick, Alberto, Waugh, and Regine (2013) indicate that regression analysis can be used as an additional, statistical analysis technique allowing researchers to gain as much information as possible concerning the effect of a treatment on the outcome scores.

6.1.1 Introduction to the regression-based approach

The regression equation that can be used to analyze a simple AB phase design looks as follows and the interpretation of the coefficients is presented in Figure 6.1:

$$Y_i = \beta_0 + \beta_1 Treatment_i + e_i \text{ with } e_i \sim N(0, \sigma_e^2) \quad (6.1)$$

In Equation 6.1, Y_i indicates the outcome score on the dependent variable at measurement occasion i (i = 0, 1,…*I*), $Treatment_i$ is a dummy coded variable that equals zero when measurement occasion i belongs to the baseline, and one otherwise. Therefore, β_0 indicates the expected baseline level and the coefficient β_1 represents the treatment effect.

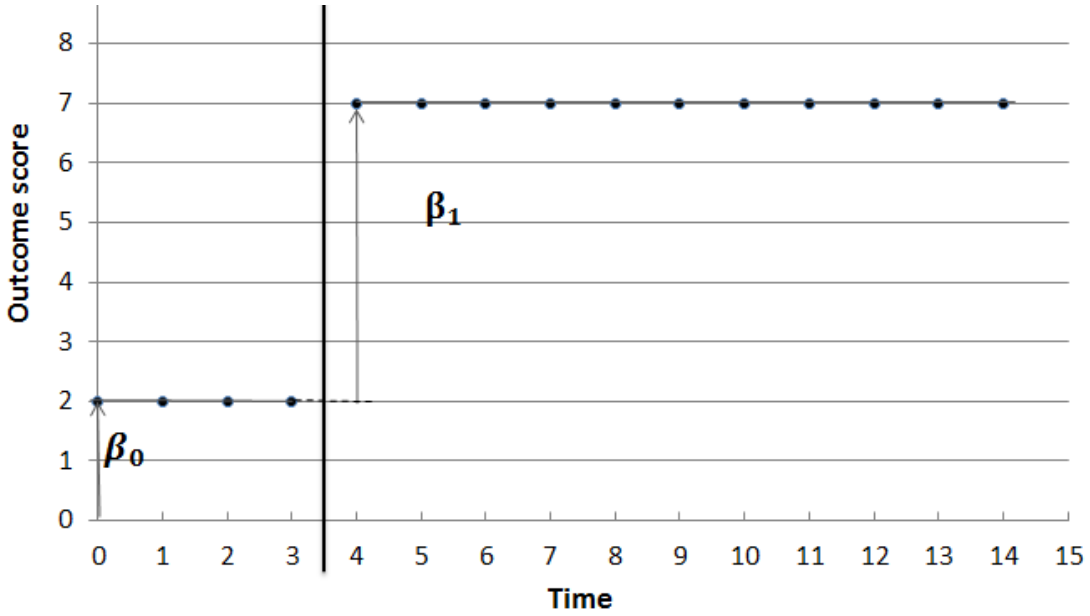

Figure 6.1 Graphical presentation of the coefficients from equation 6.1 for hypothetical data.

In SSEDs, outcome scores on some behavior are obtained across consecutive measurement occasions, and therefore we can add a time variable as suggested by Center et al. (1985-1986). In Equation 6.2, $Time_i$ is a time variable, which plays a crucial role in SSEDs, and is coded with a zero at the beginning of the baseline phase. $Treatment_i Time_i$ is a variable representing the interaction between the dummy coded treatment and time variable. Because the SSED researcher is commonly interested in the change in level (i.e., the immediate treatment effect), defined as the change between the estimated value based on the baseline phase regression and the treatment phase regression at the first measurement occasion of the treatment phase, we propose centering the $Treatment_i Time_i$ variable around the first measurement occasion of the treatment phase (see Figure 6.2). The centered time variable used in the interaction term ($Treatment_i Time_i$) is indicated by *Time1* in Figure 6.2 and is centered by subtracting the sum of one plus the number of measurement occasions, n_1, in the baseline phase {i.e., $[Time_{ij} - (n_1 + 1)]$}(Center et al., 1985-1986; Huitema & McKean, 2000). As a consequence, values on *Time*1 during the baseline phase are coded with negative values, counting downwards from the start of the treatment phase to the beginning of the baseline phase (see Figure 6.2).

$$Y_i = \beta_0 + \beta_1 Time_i + \beta_2 Treatment_{ij} + \beta_3 Treatment_i [Time_i - (n_1 + 1)] + e_i \quad (6.2)$$

$$\text{with } e_i \sim N(0, \sigma_e^2)$$

The time variable, $Time_i$, can be expressed in days, session number, etc. The interpretation of the coefficients from Equation 6.2 is displayed in Figure 6.2. In the baseline phase, the expected score at measurement occasion i, Y_i, equals $\beta_0 + \beta_1 Time_i$, while the expected score is $\beta_0 + \beta_1 Time_i + \beta_2 + \beta_3 Time1_i$ in the treatment phase. Therefore, β_0 indicates the expected baseline level at the start of the baseline phase (when $Time_i$ = 0 and $Treatment_i$ = 0) and β_1 is the linear trend over time in the baseline phase scores. The coefficient β_2 represents the difference between the predicted value of Y_i at the beginning of the treatment phase ($Time1_i$ equal zero) under the treatment phase and the predicted value at the same point in time under the baseline phase (see Figure 6.2). β_3 is the difference in slope between baseline and treatment phases (i.e., the change in slope due to the treatment). In Figure 6.2, we use β_4 to indicate the slope during the treatment phase. Although this coefficient is not presented in regression Equation 6.2, we included it in the graphical presentation in order to indicate that β_3 is the difference between the slope during the baseline phase, β_1, and the slope during the treatment phase, β_4 (i.e., $\beta_3 = \beta_4 - \beta_1$).

SSED researchers are mostly interested in β_2 and β_3, but if the research interest lies in the actual trend during the treatment it is straightforward to calculate it from β_1 and β_3 (i.e., $\beta_4 = \beta_1 + \beta_3$).

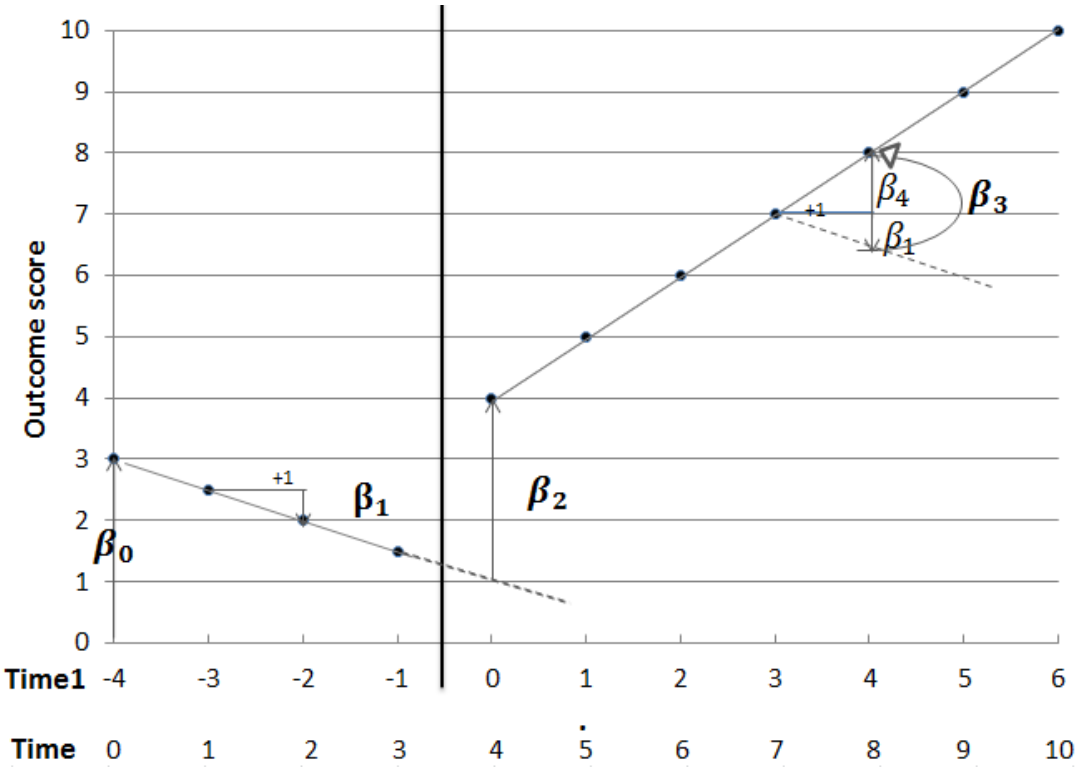

Figure 6.2. Graphical presentation of the coefficients from equation 6.2 for hypothetical data. The *X*-axis represents the variables *Time* and *Time1*. *Time* is coded such that *Time* = 0 for the first measurement occasion in the baseline phase. *Time1* is recoded such that *Time1* = 0 for the first measurement occasion in the treatment phase.

When using the regression approach to analyze SSED data, a variety of different options for coding the *Time* variable are possible and the interpretation of the coefficients depends on this coding. In previous research, attention is given to the coding of the time variable in growth curve models (Anumendem, De Fraine, Onghena, & Van Damme, 2011), but in the area of SSED research this is still underdeveloped. In this paper, we will discuss several research questions that SSED data analysts may have when analyzing results from multiple-baseline across subjects designs, reversal designs, and alternating treatment designs, because these are the three most popular SSEDs. For each particular SSED and associated research questions, we propose several design matrices and we describe and illustrate graphically how to interpret the associated parameters. According to the coding of the *Time* variable, the interpretation of the regression coefficients changes, an issue about which SSED data analysts should be aware. We illustrate each design matrix with an empirical example. The raw data for the empirical examples were retrieved from real studies using the statistical software

DataThief III (Tummers, 2005-2006). In order to assess reliability of data extraction using *DataThief III*, the extraction process was repeated twice by two independent researchers and the same raw data were obtained. The purpose of this manuscript is to guide SSED researchers and meta-analysts in the retrospective analysis of their SSED data. The quantitative analysis is complementary to the visual analysis and can provide additional information such as estimates of effect size, within-case variability, trends, and autocorrelation. Through meta-analysis, additional research questions can be examined such as: was the treatment effect consistent across subjects or is there a large difference in treatment effect between subjects?

6.1.2 Assumptions underlying the regression-based approach

When using the piecewise regression equation as presented in Equation 6.2, notice that a linear trend during both baseline and treatment phases is assumed. This might not be the case in reality and therefore there might be a need to take non-linear trajectories into account. This can be accomplished, for instance, by adding quadratic terms ($Time_i^2$) in Equation 6.2 when a quadratic trajectory is expected. It might also be a good option to simplify Equation 6.2 by removing the time trend if no linear time trend is expected. In addition, the errors are assumed to be normally, independent and identically distributed. The regression equation can be extended by modeling dependent errors, and predictors, among other complexities but this is beyond the scope of the article. For a more in depth discussion of these extensions, we refer to Moeyaert, Ferron, Beretvas, and Van den Noortgate (2014). In the remainder of this article, we will assume linear trends in the baseline and the treatment phase and that the errors are normally, identically, and independently distributed. The coding strategies illustrated, however, are general in that they could be applied with models that made different assumptions about the errors and they could be adapted to handle non-linear trends.

6.2 Analyzing Multiple-Baseline Design Data

A multiple-baseline design is one of the variants of SSEDs in which an AB phase design (with one baseline phase, A, and one treatment phase, B) is delivered simultaneously to different participants, behaviors or settings (Barlow & Hersen, 1984; Ferron & Scott, 2005; Kazdin, 2011; Onghena, 2005). The staggering of the initiation of treatment across participants, settings or behaviors, allows researchers to disentangle a change in data due the introduction of the treatment and external event effects (Baer et al., 1968; Barlow & Hersen, 1984; Kinugasa et al., 2004; Koehler & Levin, 2000).

We can choose to analyze the data for each participant separately (in a single-level analysis) or to analyze the data from multiple participants simultaneously (in a two-level analysis, see Van den Noortgate & Onghena, 2003a). We use the multiple-baseline study of Laski, Charlop, and Schreibman (1988) to illustrate both approaches. In their study, the effects of parents training in using the natural language paradigm to increase autistic children's speech were investigated for nine children. In all cases, the treatment increased the amount of spontaneous speech.

6.2.1 Single-level analysis

6.2.1.1 Design matrix 1

For the single-level analysis, we consider the multiple-baseline design as separate AB phase designs. If the SSED researcher is mainly interested in the treatment effect estimates for each participant separately and not in an average estimate over the participants, than we can simply use the piecewise regression equation proposed by Center et al. (1985-1986) and presented in Equation 6.2. The results of this analysis are displayed in Table 6.1 and the SAS code is included in Addendum A3. We only present the results for the first two subjects (see Figure 6.3) in order to reduce size of the table. The raw data are included in Addendum B.

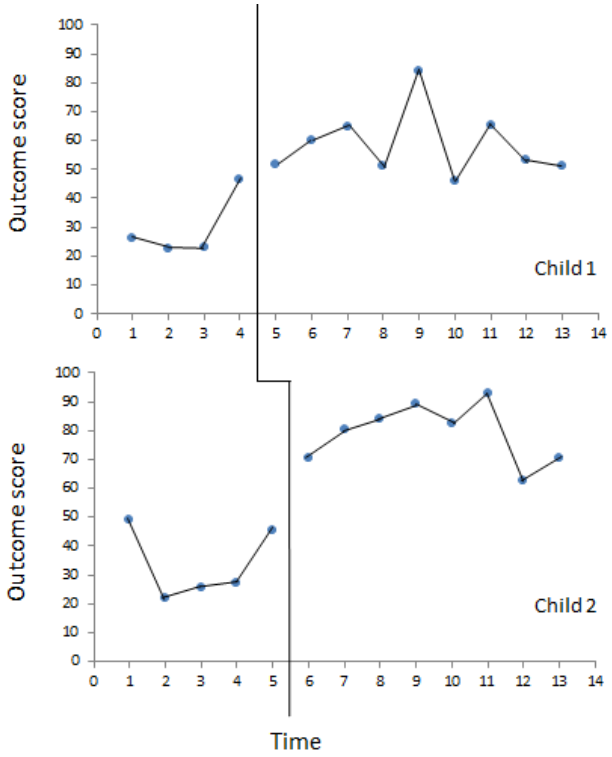

Figure 6.3. Graphical presentation of a multiple-baseline design across two participants using the data from "Training Parents to use Natural Language Paradigm to increase their Autistic Children's' speech" by Laski, K. E., Charlop, M. H., and Schreibman, L., 1988, *Journal of Applied Behavior Analysis, 21*, p.391-400.

The estimate of the predicted outcome score at the start of the baseline phase, $\hat{\beta}_0$, the slope during the baseline $\hat{\beta}_1$, the immediate treatment effect, $\hat{\beta}_2$, and the change in slope due to the treatment, $\hat{\beta}_3$, correspond to what we expect from the graphical presentation of the raw data in Figure 6.3. The outcome score at the beginning of the baseline phase differs significantly from zero for the second participant and equals 35.48, $t(9) = 3.95$, $p = .003$. The trend during the baseline is positive for both participants, but not statistically significant. For the Laski et al. (1988) data, $\hat{\beta}_2$ equals 16.51 ,$t(9) = 0.99$, $p = .35$, and 48.02, $t(9) = 3.36$, $p = .008$ for the first and the second subject, respectively. The estimated change in trend for the first participant is more than 18 times larger; $\hat{\beta}_3$ = -6.43, $t(9) = 0.03$, $p = .03$, than the estimated change in trend for the second participant; $\hat{\beta}_3$= -0.34, $t(9) = -0.28$, $p = .78$, but neither estimate is statistically significant. In this example, the interpretation of the immediate treatment effect, $\hat{\beta}_2$, represents the difference at the start of treatment (*Time1* = 0) between the estimated outcome score based on the baseline phase's regression model versus the estimated outcome score using the treatment phase's regression model (see Figure 6.2).

Design matrix 1 can be used if the research questions of interest include:
(a) What is the outcome score at the beginning of the baseline phase (β_0)?
(b) What is the trend during the baseline phase (β_1)?
(c) What is the immediate treatment effect (β_2)?
(d) What is the change in trend between the baseline phase and the treatment phase (β_3)?
Design matrix 1 is of particular interest because SSED researchers are typically interested in the immediate treatment effect (research question c) and the treatment effect on the slope (research question d).

In the following paragraphs we suggest three alternative design matrixes that can be used for the single-level analysis of multiple-baseline design data along with a graphical presentation of the coefficients' interpretation for hypothetical data. We finish with an empirical illustration using a real dataset. For the empirical illustration for each design matrix, we choose to include only the data for two participants of the study of Laski et al. (1988) instead of the data from all nine participants. However, the same single-level analysis could be conducted for the seven other participants.

6.2.1.2 Design matrix 2

It might be the case that the SSED analyst chooses to leave the *Time* variable uncentered in the $Treatment_i Time_i$ interaction term in Equation 6.2 as follows:

$$Y_i = \beta_0 + \beta_1 Time_i + \beta_2 Treatment_i + \beta_3 Treatment_i Time_i + e_i \text{ with } e_i \sim N(0, \sigma_e^2) \quad (6.3)$$

This has consequences for the interpretation of the $\hat{\beta}_2$ coefficient as it no longer represents the change in outcome scores between baseline and treatment at the start of the treatment phase, see Figure 6.4.

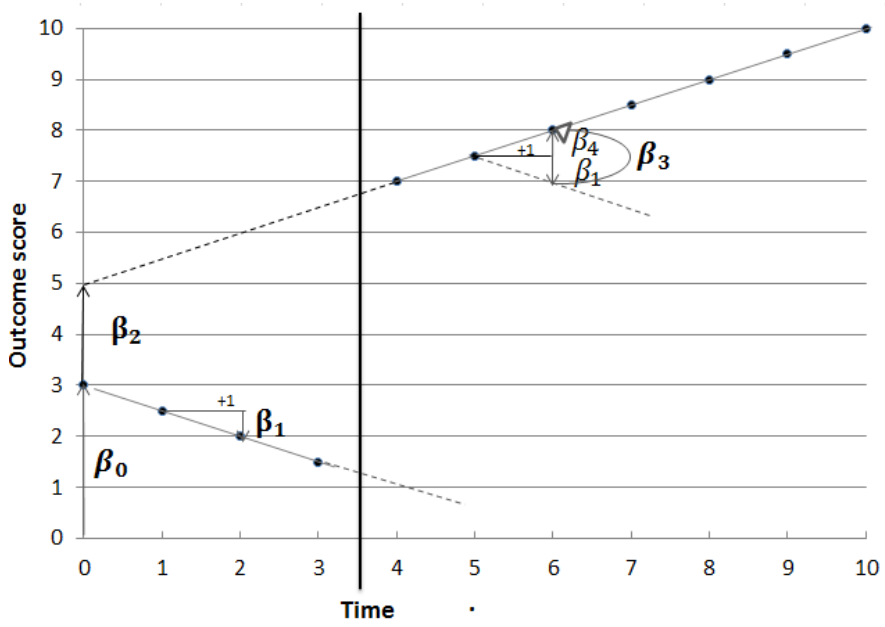

Figure 6.4. Graphical presentation of the coefficients from equation 6.3 for hypothetical data.

The interpretation of the immediate treatment effect, $\hat{\beta}_2$, using design matrix 2 now is the difference at the start of the baseline phase (*Time* = 0) between the estimated outcome score based on the baseline regression model versus the estimate using the treatment phase regression model (see Figure 6.4). This is unlikely an interesting parameterization for an SSED researcher. Interpretation of the other coefficients (including predicted outcome score at the start of the baseline phase, $\hat{\beta}_0$, the slope during the baseline phase, $\hat{\beta}_1$, and the change in slope due to the treatment, $\hat{\beta}_3$) remains the same as under the first design matrix. The results of this analysis are displayed in Table 6.1 and the SAS code that was utilized is included in Addendum A3. Nevertheless, design matrix 2 can be used if the research questions of interest are:

(a) What is the outcome score at the beginning of the baseline phase (β_0)?

(b) What is the trend during the baseline phase (β_1)?

(c) What is the change in trend between the baseline phase and the treatment phase (β_3)?

6.2.1.3 Design matrix 3

The $\hat{\beta}_2$ coefficient in design matrix 2 is likely not of interest for a typical SSED analyst and even more problematic is that the change in level cannot be estimated. Because the SSED researcher is commonly interested in the change in the outcome's level (i.e., the immediate treatment effect), defined as the change between the estimated value based on the baseline regression and the treatment regression at the first measurement occasion of the treatment phase, we propose centering the time variable used in the calculation of the interaction effect around the first measurement occasion of the treatment phase as proposed in design matrix 1. If the research interest focuses solely on the immediate treatment effect and the treatment effect on the slope, design matrix 1 can be simplified by using one centered time variable: $Time1_i$ [i.e., $Time1_i = Time_i - (n_1 + 1)$] instead of using two different time indicators (i.e., $Time_i$ and $Time1_i$ in the interaction term in design matrix 1). In addition, design matrix 3 answers another important research question: What would to outcome value have been at the start of the treatment phase, had the baseline phase continued? Baseline observations are used to document the need for an intervention and therefore it may be more interested to document how problematic the measured dependent variable was at the time of intervention ($\hat{\beta}_0$ from design matrix 3) than to know how problematic it was at some earlier point in time ($\hat{\beta}_0$ from design matrix 1). This results in the following:

$$Y_i = \beta_0 + \beta_1[T_i - (n_1 + 1)] + \beta_2 Treatment_i + \beta_3 Treatment_i[\, T_i - (n_1 + 1)] + e_i \quad (6.4)$$
$$\text{with } e_i \sim N(0, \sigma_e^2)$$

The results of this analysis are displayed in Table 6.1 and the necessary SAS code is included in Addendum A3. The only difference between design matrix 1 and design matrix 3 lies in the interpretation of $\hat{\beta}_0$ (see Figure 6.5).

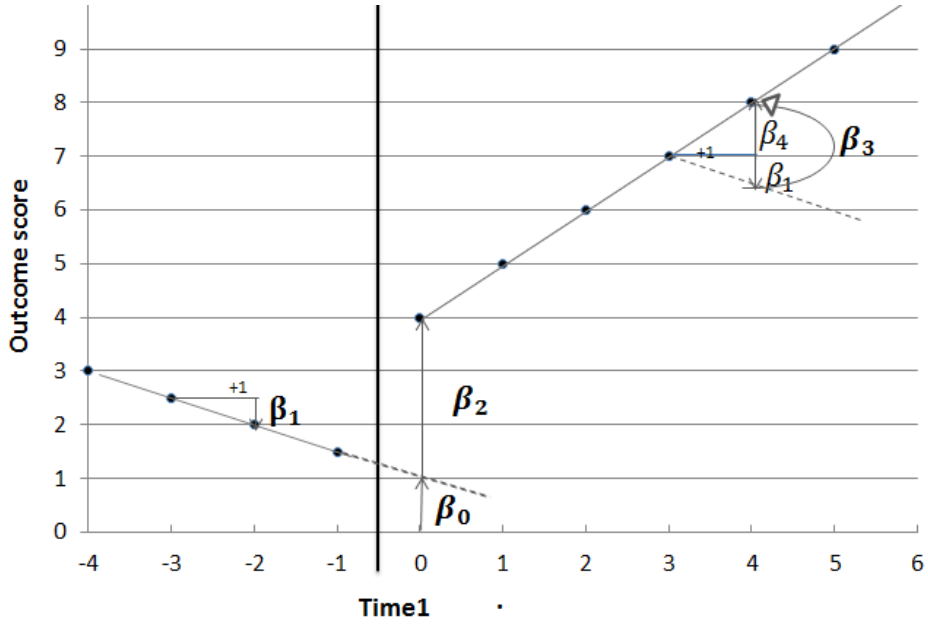

Figure 6.5. Graphical presentation of the coefficients from equation 6.4 for hypothetical data. The *X*-axis represents the variable *Time1* which is recoded such that *Time1* = 0 for the first measurement occasion in the treatment phase.

Using design matrix 3, $\hat{\beta}_0$ is the estimated outcome score at the beginning of the treatment phase using the data from the baseline phase and equals 45.38, $t(9) = 3.05$, $p = .014$, and 33.31, $t(9) = 2.74$, $p = .02$, for participant 1 and 2 respectively (see Table 6.1). This means that $\hat{\beta}_0$ is statistically significant and its value is larger in comparison to estimates when using the previous proposed design matrices. Moreover, when an SSED researcher chooses to use design matrix 3, they should be aware that the interpretation of the intercept changes depending on the length of the baseline phase (which itself determines the timing of the first intervention measurement occasion).

6.2.1.4 Design matrix 4

In design matrices 1 through 3, the β_3 coefficient indicated the difference in slope during the baseline versus treatment phases. An SSED researcher might not be interested in this change, but rather in the value of the slope during the treatment phase, β_4 (i.e., the change in slopes can still be calculated as the difference in the estimated slopes in both phases as discussed in design matrix 1). If this is the case, we propose to set the time variable (now represented by *Time2* in Figure 6.6) to a constant value during the treatment phase representing the time variable's value at the first measurement occasion during intervention (here, a value of four, see the values on the *X*-axis in Figure 6.6).

Using design matrix 4, the trend during the treatment, β_4, is obtained directly with its standard error and *p*-value, and the interpretation of the other coefficients, β_0, β_1, and β_2 remains the same as in design matrix 2 (see Table 6.1). Equation 6.4 can be used to estimate the coefficients of interest:

$$Y_i = \beta_0 + \beta_1 Time2_i + \beta_2 Treatment_i + \beta_4 Treatment_i[\,T_i - (n_1 + 1)] + e_i \quad (6.5)$$

$$\text{with } e_i \sim N(0, \sigma_e^2)$$

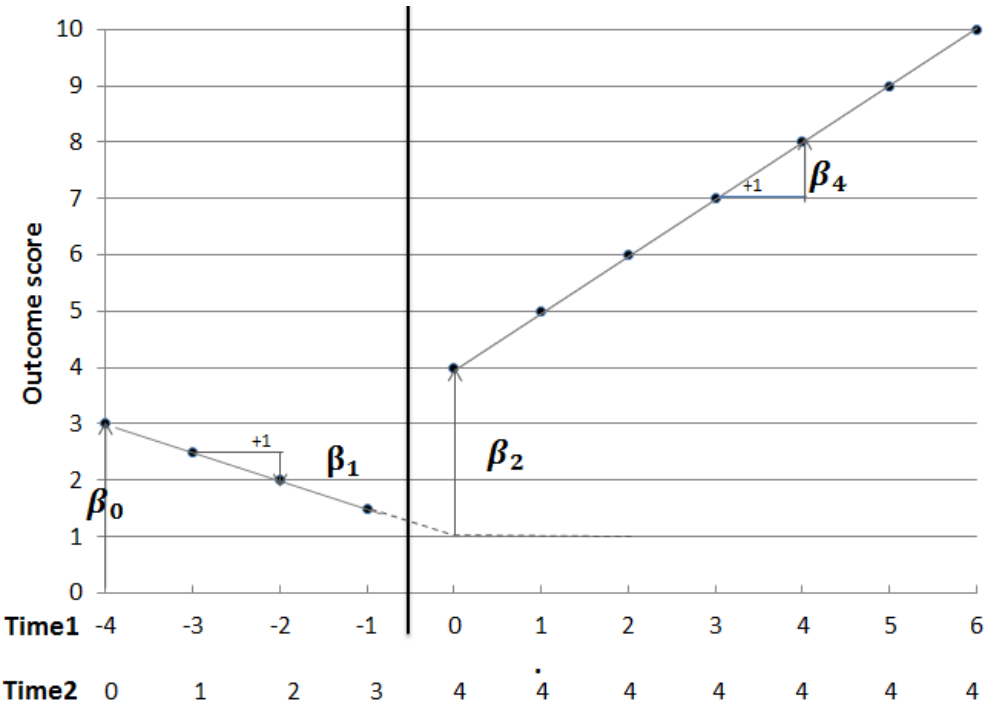

Figure 6.6. Graphical presentation of the coefficients from Equation 6.5 for hypothetical data. The *X*-axis represents the variable Time1 which is recoded such that *Time1* = 0 for the first measurement occasion in the treatment phase.

The results of this analysis are displayed in Table 6.1 and the SAS code that was used is included in Addendum A3.

Using this design matrix, the following research questions can be evaluated:

(a) What is the outcome value at the beginning of the baseline phase (β_0)?

(b) What is the trend during the baseline phase (β_1)?

(c) What is the immediate treatment effect (β_2)?

(d) What is the trend during the treatment phase (β_4)?

Using design matrix 4, we found that the linear slope for participant 1 equals 5.86, $t(9)$ = 5.43, p = .31, during baseline and -0.55, $t(9)$ = 1.57, p = .74, during treatment, this corresponds to a change in slope of -6.11 (-0.55 - 5.86), and equals well approximated estimates of β_3 using design matrices 1 through 3; β_3 = -6.43, $t(9)$ = -1.14, p = .28. This was expected because β_3 in design matrices 1 through 3 refers to the difference in slopes between the baseline and the treatment phase. A similar result was found for participant 2: $\hat{\beta}_3 - \hat{\beta}_1 = -0.77 - (-0.43)$ = -0.34 and this equals the β_3 estimate that was found when using matrices 1 through 3; -0.34, $t(9)$ = 0.08, p = .94.

6.2.1.5 Conclusion - single-level analysis of multiple baseline design

The results of the single-level analysis using four different design matrices applied to the first two participants of the study of Laksi (1988) are presented in Table 6.1. Design matrices 1 through 3 build further upon the design matrices proposed by Huitema and McKean (2000). In summary, we advocate most strongly for using design matrix 1 (in which the *Time* variable in the interaction term is centered around the intervention phase's starting point), because it is more likely that an SSED researcher is interested in the change in level (immediate treatment effect) at the first treatment phase measurement occasion and the change in slope due to the treatment. Design matrix 1 also entails the advantage that the outcome score at the beginning of the baseline phase can be estimated. If one is not interested in the outcome score at the beginning of the baseline phase, design matrix 3 can instead be used. Design matrix 4 is also of practical use if the research interest lies in the estimate of the slope during the treatment instead of in the change in slope due to the treatment. And note that the slope during the treatment can also be calculated indirectly via the estimate of the change in slope between baseline and treatment phase, β_3, and the slope during the baseline phase, β_1. A drawback of this latter approach is that the standard errors and p-values are not directly obtained and have to be calculated by hand.

Table 6.1

Summary Results Single-Level Regression Analysis: Design Matrix 1 – Design Matrix 4

Coefficient		Matrix 1 Estimate (*SE*)	Matrix 2 Estimate (*SE*)	Matrix 3 Estimate (*SE*)	Matrix 4 Estimate (*SE*)
Participant 1					
Score when *Treatment*=0 and *Time*= 0	$\hat{\beta}_0$ [a]	21.84 (10.16)	21.84 (10.16)	45.38*(14.88)	21.84 (10.16)
Linear trend during the baseline	$\hat{\beta}_1$	5.86 (5.43)	5.86 (5.43)	5.86 (5.43)	5.86 (5.43)
Treatment effect on level	$\hat{\beta}_2$ [b]	16.51 (16.64)	42.23*(16.64)	16.51 (16.64)	16.51 (16.64)
Treatment effect on slope	$\hat{\beta}_3$ [c]	-6.43 (5.65)	-6.43 (5.65)	-6.43 (5.65)	-6.11 (5.65)
Participant 2					
Score when *Treatment*=0 and *Time*= 0	$\hat{\beta}_0$ [a]	35.48* (8.98)	35.48* (8.98)	33.31*(12.16)	35.48* (8.98)
Linear trend during the baseline	$\hat{\beta}_1$	-0.43 (3.67)	-0.43 (3.67)	-0.43 (3.67)	-0.43 (3.67)
Treatment effect on level	$\hat{\beta}_2$ [b]	48.02*(14.28)	49.72*(18.13)	48.92*(14.28)	48.02* (14.28)
Treatment effect on slope	$\hat{\beta}_3$ [c]	-0.34 (4.08)	-0.34 (4.08)	-0.34 (4.08)	-0.34 (4.08)

Note. [a]$\hat{\beta}_0$ represents the outcome score at the start of the baseline in design matrix 1, design matrix 2 and design matrix 4. In design matrix 3, $\hat{\beta}_0$ represents the outcome score at the start of the treatment using baseline data. [b]$\hat{\beta}_2$ indicates the immediate treatment effect in design matrix 1, design matrix 3 and design matrix 4. In design matrix 2, $\hat{\beta}_2$ refers to the treatment effect on the first measurement occasion in the baseline phase. [c]$\hat{\beta}_3$ is the treatment effect on the slope in design matrix 1, design matrix 2 and design matrix 3. In design matrix 4, the treatment effect on the slope is calculated indirectly, using the estimated slope during the baseline, $\hat{\beta}_1$, and the estimated slope during the treatment, $\hat{\beta}_4$.
* $p < .05$.

6.2.2 Two-level analysis

It might be time consuming to analyze data from a multiple-baseline design for each subject separately. There is an increased interest in using scaled-up multiple-baseline designs. For instance, the multiple-baseline design study of Koutsoftas, Harmon, and Gray (2009) included 36 participants. Therefore we propose using a two-level analysis which allows estimating the treatment effect across participant without losing information about the individual participants (Van den Noortgate & Onghena, 2003a). Moreover, this analysis takes into account that measurement occasions are nested within participants and it allows modeling of and estimating the variability in these treatment effects as well. We will illustrate the two-level analysis using design matrix 1 (i.e., the *Time* variable within the interaction term is centered around the first measurement occasion of the treatment). Note that the analysis is similar if using any of the other design matrices described earlier. We chose design matrix 1, because this design matrix allows estimation of (1) the average outcome score at the beginning of the baseline phase, (2) the average trend during the baseline phase, (3) the

average immediate treatment effect, and (4) the average treatment effect on the trend across participants. Also the between-case variability in each of these four parameter estimates can be estimated. We can extend Equation 6.2 from the single-level analysis by adding an additional index, *j*, indicating the subject. At the first level of the two-level model, measurement occasions, *i*, are nested within participants, *j*. In contrast to the single-level model, we allow the coefficients from the first level, β_{0j}, β_{1j}, β_{2j} and β_{3j} to vary at the second (participant) level:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}Time_{ij} + \beta_{2j}Treatment_{ij} + \beta_{3j}Treatment_{ij}[Time_{ij} - (n_1 + 1) + e_{ij} \quad (6.6)$$

$$\text{with } e_{ij} \sim N(0, \sigma_e^2)$$

Level 2:

$$\begin{cases} \beta_{0j} = \theta_{00} + u_{0j} \\ \beta_{1j} = \theta_{10} + u_{1j} \\ \beta_{2j} = \theta_{20} + u_{2j} \\ \beta_{3j} = \theta_{30} + u_{3j} \end{cases} \text{ with } \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim N(0, \Sigma_u) \quad (6.7)$$

The first equation in Equation 6.7 indicates that the baseline intercept for participant *j* equals an average baseline intercept, θ_{00}, plus a random deviation from this mean, u_{0j}. The subsequent equations describe the variation across participants from the same study in the time effect in the baseline phase, the immediate treatment effect, and the treatment effect on the linear trend, respectively. SSED researchers are especially interested in the immediate treatment effect across participants, θ_{20}, the treatment effect on the time trend across participants, θ_{30}, the between-case variability of the immediate treatment effect, $\sigma^2_{u_{2j}}$, and the between-case variance of the treatment effect on the time trend, $\sigma^2_{u_{3j}}$. The results of the two-level analysis applied to the nine participants of the study of Laski et al. (1988) are displayed in Table 6.2 and the SAS code that was used to estimate the parameters is included in Addendum A4.

The outcome score predicted at the beginning of the baseline phase across the nine participants ($\hat{\theta}_{00}$) equals 37.72, $t(1.23) = 4.33$, $p = .002$, and is statistically significant. There is a small average downward linear time trend during the baseline phase, $\hat{\theta}_{10}$ = -0.15, $t(34.8)$ = -0.24, $p = .81$, and the estimated average immediate treatment effect ($\hat{\theta}_{20}$) equals 32.10, $t(6.88) = 5.52$, $p < .001$. This means that the treatment has an immediate, significant positive effect on speech across the nine participants. The change in slope due to the treatment equals

0.70, $t(43.7) = 0.81$, $p = .42$. Despite obtaining an average estimate of the effects over participants, participant-specific effects can also be obtained using empirical Bayes estimates. To obtain these estimates in the output, we simply add the command: "solution" after the random statement in the PROC MIXED procedure detailed in Addendum A4. This two-level approach also allows estimation of the between-case variability of the outcome score at the start of the baseline phase ($\sigma^2_{u_{oj}}$), the average trend during the baseline ($\sigma^2_{u_{1j}}$), the average immediate treatment effect ($\sigma^2_{u_{2j}}$), and the average treatment effect on the time trend ($\sigma^2_{u_{3j}}$), and within-case variability ($\sigma^2_{e_{ij}}$). For the dataset being analyzed, here, the following results were obtained: significant between-case variability in terms of the outcome score at the start of the baseline; $\sigma^2_{u_{oj}} = 520.59$, $Z = 1.92$, $p = .027$, and the within-case variability; $\sigma^2_{e_{ij}} = 148.67$, $Z = 7.54$, $p < .0001$. None of the between-case variances (in the trend during the baseline, the immediate treatment effect and the treatment effect on the time trend) differed significantly from zero.

Table 6.2

Two- Level Analysis using Design Matrix 2 Applied to the Study of Laski et al. (1988)

Coefficient	Parameter	Estimate	(*SE*)
Fixed coefficient			
Outcome score at the start of the baseline	$\hat{\theta}_{00}$	37.72*	(8.01)
Trend during the baseline	$\hat{\theta}_{10}$	-0.15	(0.64)
Immediate treatment effect	$\hat{\theta}_{20}$	32.10*	(5.82)
Treatment effect on the slope	$\hat{\theta}_{20}$	0.70	(0.86)
Between-case variance			
Outcome score at the start of the baseline	$\hat{\sigma}^2_{u_0}$	520.59*	(271.11)
Trend during the baseline	$\hat{\sigma}^2_{u_1}$	-	-
Immediate treatment effect	$\hat{\sigma}^2_{u_2}$	162.42	(105.30)
Treatment effect on the slope	$\hat{\sigma}^2_{u_3}$	0.18	(1.57)
Residual within-case variance	$\hat{\sigma}^2_e$	148.67*	(19.72)

Note. * $p < .05$.

6.2.2.1 Conclusions - two-level analysis of multiple-baseline designs

As with the single-level model's parameterization, for the two-level analysis of multiple-baseline designs we also encourage use of the design matrix in which the *Time* variable is centered around the start of the intervention phase. This enables researchers to estimate the between-case variance in the immediate treatment effect and the treatment effect on the time trend at the first measurement occasion of the treatment phase. Moreover by only centering the time variable involved in the interaction term, we can estimate the between-case variance in the outcome score at the beginning of the baseline phase. A possible problem when centering the time variable and the time variable in the interaction term as in design matrix 3 is that if there is variation over cases in the baseline trend, the between case-variance depends on the measurement occasion around which the time variable is centered. As a consequence, if the length of the baseline varies over cases (as is common in a MBD), the assumption of homoscedasticity might be violated. The two-level model is more efficient and provides more information in comparison to the use of level-1 regression equation for each participant. More specifically, we can estimate the effects (e.g. immediate treatment effect and treatment effect on time trend) across participants which allow us to make more general conclusions. We can also estimate the participant-specific effects using empirical Bayes estimates. Using this two-level model allows estimation of within- and between-case variability.

6.3 Reversal Designs

Reversal designs, including for instance the ABAB design involve introduction and withdrawal of the treatment. In these kinds of designs, there is more than one baseline and treatment phase per participant.

We will divide this section into two parts. In the first part, we propose a design matrix to use when the researcher is interested in the average outcome score in the baseline, the average treatment effect across the phases, and in the difference between the treatment effect estimates between the two AB pairs. In the second part, we are interested in the differences in outcome score and trends between consecutive phases and propose three different design matrices. In both parts, we present the regression equation(s) that can be used to analyze some hypothetical data, and a graphical representation to help understand the coefficients. We end using data from Moes' (1998) study to provide an empirical illustration. In Moes' study, the author evaluated the effects of choice making (students vs. tutor) on challenging behavior of

four children with autism using an ABAB design. The results indicated (based on visual analysis) a reduction in challenging behaviors during the treatment phase. We randomly selected the data for one of the students to illustrate the proposed designs matrices (see Figure 6.7).

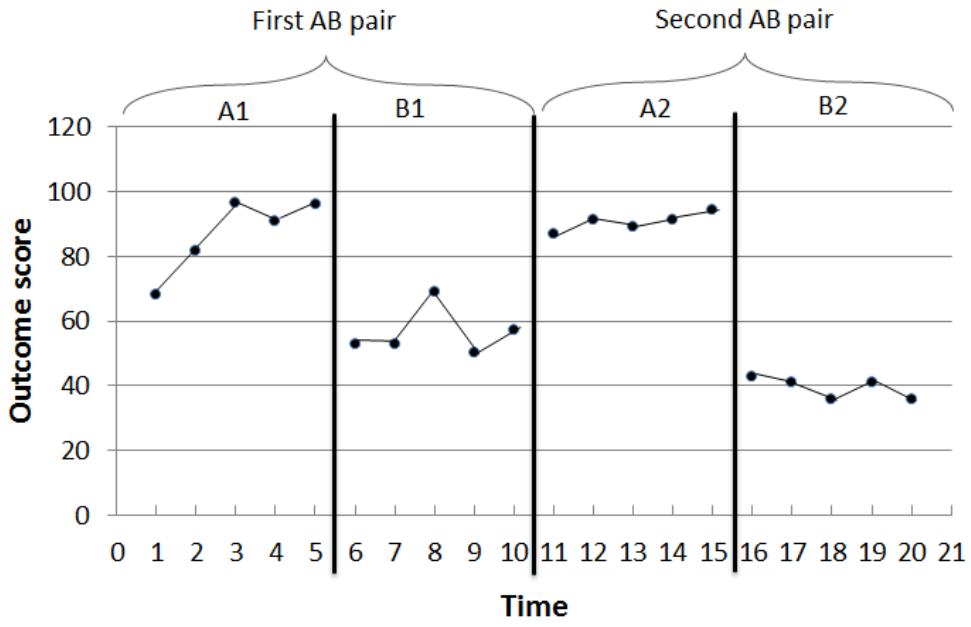

Figure 6.7. ABAB reversal design data for one participant of the study of Moes (1998).

6.3.1 First way to code phase and time in an ABAB reversal design

6.3.1.1 Design matrix 5

If a researcher wants to study change in response patterns between baseline and treatment phases (between the first and second AB pair), and wants to look at the difference in these effects between AB pairs, we can use Equation 6.8:

$$Y_{ij} = \beta_0 + \beta_1 Treatment_i + \beta_2 Pair_i + \beta_3 Treatment_i Pair_i + e_i \quad (6.8)$$

$$\text{with } e_i \sim N(0, \sigma_e^2)$$

The dummy variable $Treatment_i$ indicates if measurement i is part of the baseline phase (i.e., A1 or A2) or the treatment phase (i.e., B1 or B2). If the measurement occasion belongs to B1 or B2, then $Treatment_i$ equals one, otherwise zero. The dummy variable, $Pair_i$, indicates whether the measurements belong to the first or the second AB pair. $Pair_i$ equals 1 if the measurement occasion belongs to the second AB pair, zero otherwise. Using Equation 6.8, β_0 and β_1 equal the baseline level and the immediate treatment effect, respectively, during the first AB pair; $\beta_0 + \beta_2$ indicates the second baseline level and $\beta_1 + \beta_3$ refers to the

change in level in the second AB pair. In this way, β_2 indicates the change in outcome score between the second and the first baseline phase and β_3 indicates the difference in treatment effect between the first AB pair and the second AB pair. The graphical presentation of these coefficients is presented in Figure 6.8 and the results using this coding for the one participant randomly selected from the study of Moes (1998) are presented in Table 6.3.

The research questions that can be assessed using this design matrix include:

(a) Is the outcome score during the second baseline phase different than the outcome score during the first baseline phase (β_2)?

(b) Is the outcome score during the first treatment phase different than the outcome score during the first baseline phase (β_1)?

(c) Is the change in outcome score during the first AB pair different than the change in outcome score during the second AB pair (β_3)?

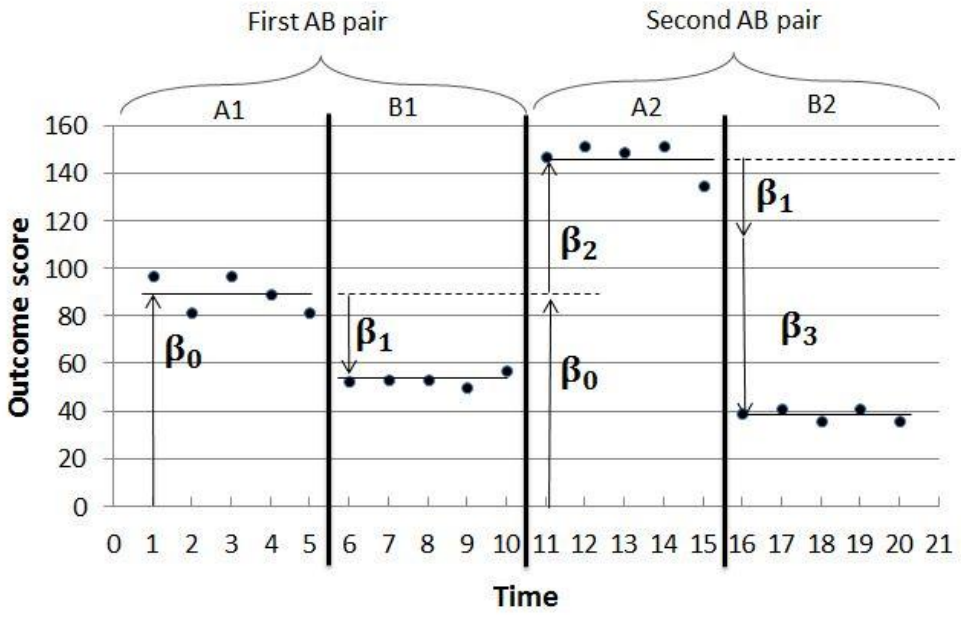

Figure 6.8. Graphical presentation of the coefficients from equation 6.8 for hypothetical data for an ABAB design.

Table 6.3

Regression Analysis of an ABAB Design using the First Way of Coding Applied to the Dataset of Moes (1998)

Coefficient	Parameter	Estimate (*SE*)
Outcome at the start of the baseline	$\hat{\beta}_0$	86.71* (3.31)
Immediate treatment effect first AB pair	$\hat{\beta}_1$	-30.40* (4.69)
Difference between the outcome score during A2 and A1	$\hat{\beta}_2$	3.94 (4.69)
Difference between the immediate treatment effect during the second AB pair and the first AB pair.	$\hat{\beta}_3$	-21.08* (6.63)

Note. * $p < .05$.

From these results we can deduce that the difference in outcome during the first and the second baseline equals 3.94 and is not statistically significant, $t(1) = -6.47$, $p < .0001$. A striking and interesting result is that the immediate treatment effect during the second AB pair is significantly larger than the immediate treatment effect during the first AB pair: $\hat{\beta}_3 = -21.08$, $t(1) = -3.18$, $p = .006$.

6.3.2 Alternative way of coding an ABAB reversal design

In this second part, the SSED researcher is interested in the following research question: Is there a change in response patterns (immediate treatment effects and changes in trends) between the four phases? The design matrix is more complex and involves coding of multiple dummy coded variables indicating the phase to which a measurement occasion belongs. We choose the following notation to distinguish between the consecutive phases: A1 and A2 indicate respectively the first and the second baseline phase and B1 and B2 the first and the second treatment phase (see Figure 6.8).

For the ABAB phase design, three dummy variables, *A1B1*, *B1A2* and *A2B2* are coded (se Figure 6.9) as suggested by Shadish et al. (2013). *A1B1* =1 for all the measurement occasions after the first baseline phase, *B1A2* = 1 for all the measurement occasions after the first treatment phase and *A2B2* equals 1 during the last treatment phase. When all three dummy coded variables equal zero (i.e., *A1B1* = *B1A2* = *A2B2* = 0), then the indicated phase is the first baseline phase. Each dummy variable represents the jump from an earlier to its adjacent phase. Thus, for example, *B1A2* refers to the jump from B1 to A2. Besides these dummy coded variables, we suggest coding multiple *Time* variables in order to estimate changes in trends. The way the *Time* variables are coded is dependent on the research question one might have.

Research questions in which SSED researchers are likely interested when using ABAB reversal designs involve but are not limited to:

- What is the difference in trend during the first treatment phase and the second treatment phase and what is the difference in level if we jump from one phase to another (design matrix 6)?
- Has the treatment more influence on the slope during the first AB than during the second AB phase pairs? The researcher might also be interested in estimating the difference in level between consecutive phases (Design matrix 7).
- What is the change in slope between the two baseline phases and the two treatment phases and what is the difference in level between consecutive phases (design matrix 8)?

We propose three design matrices in order to directly investigate these research questions and will use the data from Moes' (1998) study to illustrate them.

6.3.2.1 Design matrix 6

In this scenario the SSED researcher is interested in:

(a) What is the difference in level between consecutive phases?

(b) What is the difference in trend between the first treatment phase and the second treatment phase?

In addition to the three dummy coded variables indicating the phase, we need to code four time variables (see *Time, Time1, Time2*, and *Time3* in Figure 6.9). The first time variable, *Time*, equals zero at the start of the first baseline phase (A1) and remains constant during the other phases. *Time1* is centered around the start of the first treatment phase (B1) and remains constant during the second baseline phase, however, it increases again during the second treatment phase. The reason for this is that we want to estimate the difference in trend between the first and the second treatment. *Time2* is centered around the first measurement occasion of the second baseline phase (A2) and is then held constant during the second treatment. *Time3* is centered around the first measurement occasion of the second treatment (B2) phase. In Figure 6.9 the coding scheme is indicated.

Y	A1B1	B1A2	A2B2	Time	Time1	Time2	Time3
68.09	0	0	0	0	-5	-10	-15
81.7	0	0	0	1	-4	-9	-14
96.59	0	0	0	2	-3	-8	-13
90.96	0	0	0	3	-2	-7	-12
96.23	0	0	0	4	-1	-6	-11
52.78	1	0	0	5	0	-5	-10
52.94	1	0	0	5	1	-4	-9
69.11	1	0	0	5	2	-3	-8
50.01	1	0	0	5	3	-2	-7
57.21	1	0	0	5	4	-1	-6
86.84	1	1	0	5	5	0	-5
91.46	1	1	0	5	5	1	-4
89.04	1	1	0	5	5	2	-3
91.46	1	1	0	5	5	3	-2
94.46	1	1	0	5	5	4	-1
42.69	1	1	1	5	6	5	0
40.89	1	1	1	5	7	5	1
35.93	1	1	1	5	8	5	2
40.89	1	1	1	5	9	5	3
35.93	1	1	1	5	10	5	4

Figure 6.9. Coding scheme for the reversal ABAB single-case design using design matrix 6.

The extension of the Center et al. (1985-1986) equation in order to estimate the parameters of interest is as follows:

$$Y_i = (\beta_0 + \beta_1 Time_i) + (\beta_2 + \beta_3 Time1_i)A1B1_i + (\beta_4 + \beta_5 Time2_i)B1A2_i + (\beta_6 + \beta_7 Time3_i)A2B2_i + e_i \text{ with } e_i \sim N(0, \sigma_e^2) \quad (6.9)$$

The coefficients from Equation 6.9 are graphically presented in Figure 6.10. Also the interpretation and the estimation of the coefficients of interest are given in Table 6.4. Similar to the graphical presentation of the change in trend in design matrix 1 to design matrix 4 in the discussion of the multiple-baseline designs, we add a coefficient, here β_8, representing the trend during the second treatment phase. This is to indicate that β_7 is the change between the slope during the first treatment phase, β_3, and the slope during the second treatment phase, β_8 (i.e., $\beta_7 = \beta_8 - \beta_3$). If an SSED researcher is interested in the trend during the treatment as well, β_8, can be easily calculated by adding β_7 to β_3. If we analyze the dataset of Moes (1998), then we obtain the results displayed in Table 6.4. Also the interpretation of the coefficients from Equation 6.9 is given in Table 6.4.

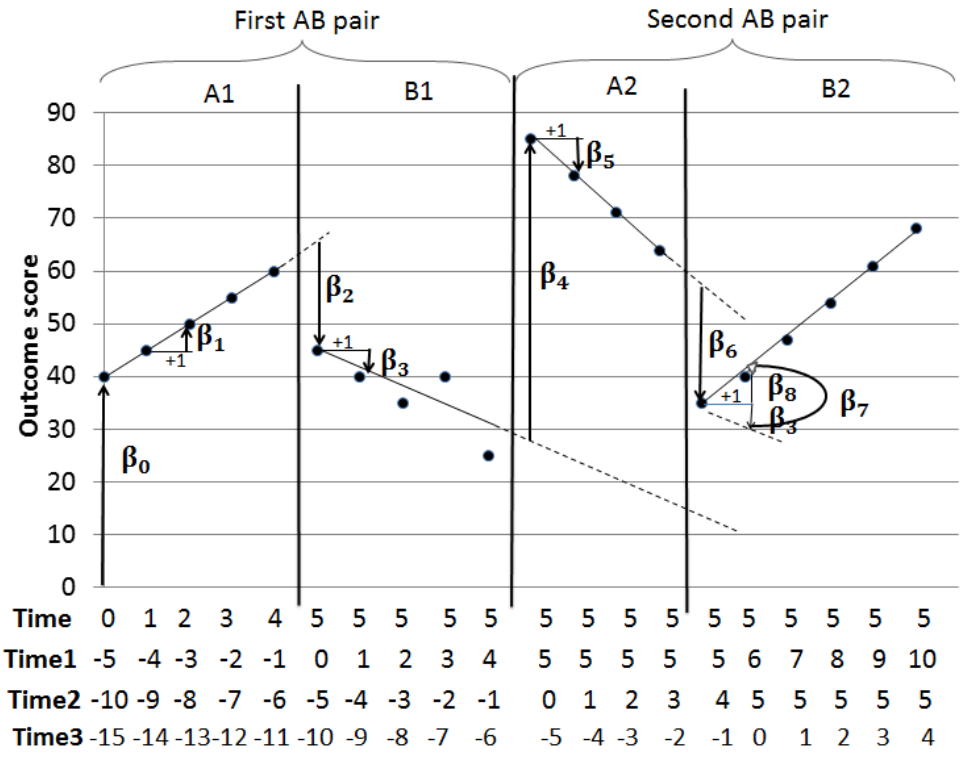

Figure 6.10. Graphical presentation of the coefficients from Equation 6.9 for hypothetical data for an ABAB design. The *X*-axis represents the variables *Time*, *Time1*, *Time2*, and *Time3*. *Time* is recoded such that *Time* = 0 for the first measurement occasion in A1, and is kept constant in consecutive phases. *Time1* is recoded such that *Time1* = 0 at the first measurement occasion in B1, is kept constant in A2, and counts further in B2. *Time2* is recoded such that *Time2* = 0 at the first measurement occasion in A2, and is kept constant in B2. *Time3* is recoded such that *Time3* = 0 at the first measurement occasion of B2.

6.3.2.2 Design matrix 7

In this scenario, the SSED researcher is interested in:

(a) What is the difference in level between consecutive phases?

(b) Has the treatment effect had more or less influence on the slope during the first AB pair in comparison to the second AB pair. Again, in addition to the three dummy coded variables indicating the phase (*A1B1*, *B1A2* and *A2B2*), we code four time variables (*Time*, *Time1*, *Time2*, and *Time3* in Figure 6.11) which are slightly different than those used in design matrix 6. *Time* is a time variable set to zero at the first measurement occasion in A1 and increasing across the phase but then remaining constant after phase B1. *Time1* is centered around the start of phase B1 and remains constant after this phase. *Time2* is centered around the start of phase A2 and *Time3* is centered around the start of phase B2. Coding time this way makes it also possible to estimate whether the change in slope during the first AB pair differs from the change in slope during the second AB pair. The results of the analysis are graphically presented in Figure 6.11. If we analyze the dataset of Moes (1998), then we obtain the results and interpretations displayed in Table 6.4. In the previous design, β_8, representing the slope

during the second treatment was inserted. In addition to β_8, another coefficient is included, namely β_9, indicating the slope during the first treatment phase. This is to indicate that β_3 is the change between the slope during the first baseline phase, β_1, and the slope during the first treatment phase, β_9.

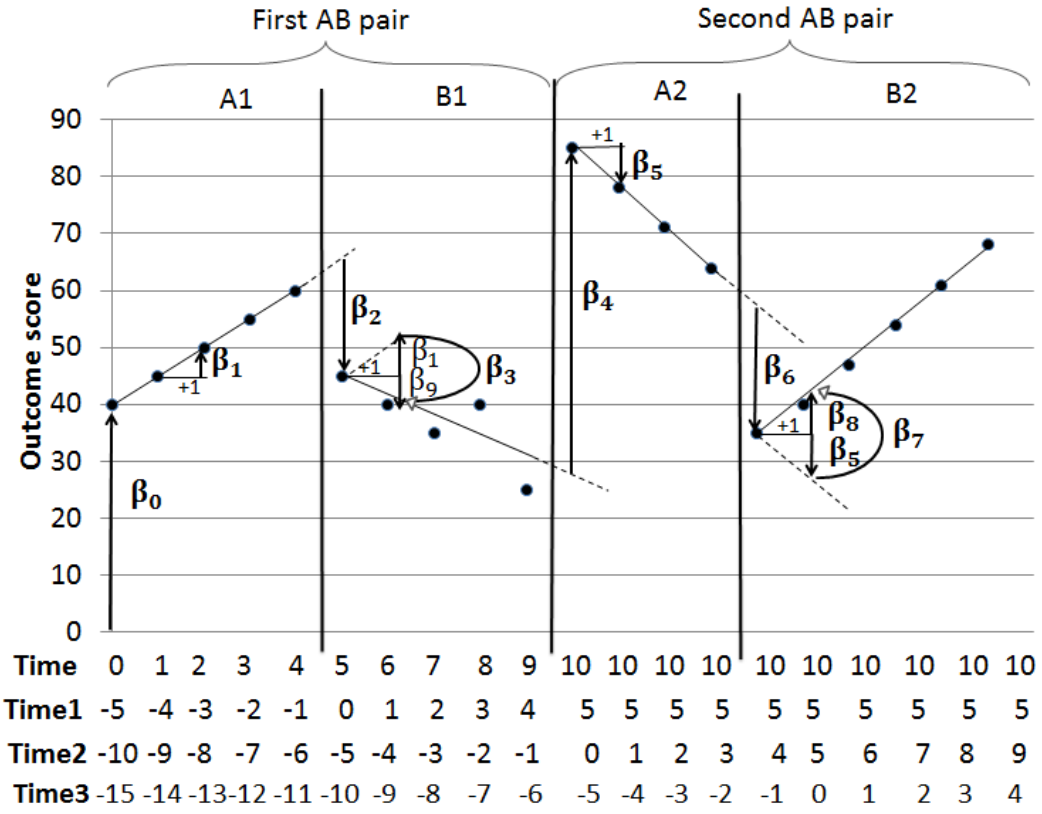

Figure 6.11. Graphical presentation of the coefficients using design matrix 7 for hypothetical data for an ABAB design. The *X*-axis represents the variables *Time*, *Time1*, *Time2*, and *Time3*. *Time* is recoded such that *Time* = 0 for the first measurement occasion in A1, and is kept constant in A2 and B2. *Time1* is recoded such that *Time1* = 0 at the first measurement occasion in B1, and is kept constant in consecutive phases. *Time2* is recoded such that *Time2* = 0 at the first measurement occasion in A2. *Time3* is recoded such that *Time3* = 0 at the first measurement occasion in B2.

6.3.2.3 Design matrix 8

In this scenario, the SSED researcher is interested in:

(a) What is the difference in level between consecutive phases?

(b) What is the difference in slope between common phases (i.e., what is the difference in the slope between A1 and A2? And what is the difference in the slope between B1 and B2)? Again we code the three dummy variables (*A1B1*, *B1A2* and *A2B2*) indicating the treatment phase as given in Figure 6.12. In this design matrix, we code four time variables slightly differently than for design matrices 6 and 7 (see *Time*, *Time1*, *Time2*, and *Time3* in Figure 6.12). *Time* equals zero at the start of phase A1. During phase B1, *Time* remains constant. During the A2 phase, *Time* begins to increase again until the start of the B2 phase. This is

because we want to compare the trend during the A2 phase with the trend during A1.We use a similar coding scheme for *Time1*. *Time1* is centered around the start of phase B1, remains constant during phase A2 and continues to increase during the B2 phase. This is because we want to compare the trend during B1 with the trend during B2. *Time2* is centered around the start of the second baseline and is set constant during the second treatment. *Time3* is centered around the start of phase A2. If we analyze the dataset of Moes (1998), then we obtain the results as displayed in Table 6.4. The coefficients of interest are graphically presented in Figure 6.12. In the previous two designs, β_8, representing the slope during the second treatment was inserted. In addition to β_8, another coefficient is included, namely β_{10}, indicating the slope during the second baseline phase. This is to indicate that β_5 is the change between the slope during the first baseline phase, β_1, and the slope during the second baseline phase, β_{10}.

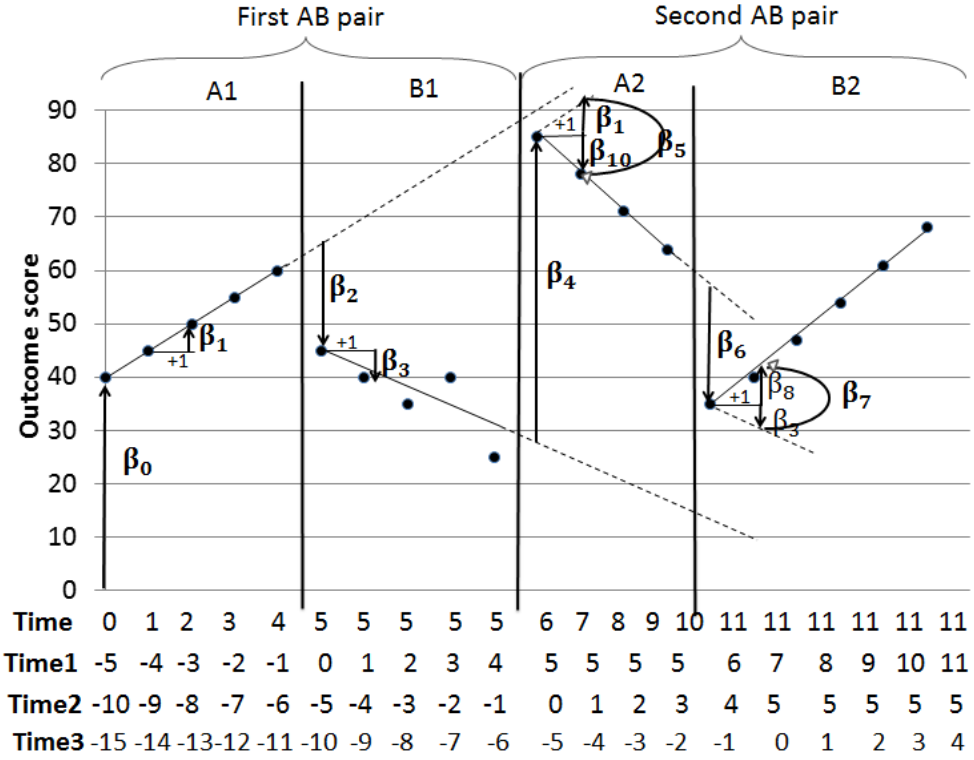

Figure 6.12. Graphical presentation of the coefficients using design matrix 8 for hypothetical data for an ABAB design. The *X*-axis represents the variables *Time*, *Time1*, *Time2*, and *Time3*. *Time* is recoded such that *Time* = 0 for the first measurement occasion in A1, is kept constant in B1, counts further in A2 and is kept constant in B2. *Time1* is recoded such that *Time1* = 0 at the first measurement occasion in B1, is kept constant in A2 and counts further in B2. *Time2* is recoded such that *Time2* = 0 at the first measurement occasion in A2 and is kept constant in B2. *Time3* is recoded such that Time3 = 0 at the first measurement occasion in B2.

6.3.3 Conclusion - reversal designs

Depending on the research questions of interest to the SSED researcher, we suggest two complementary ways of coding the design matrix for ABAB reversal designs. Under the first coding scheme, the research interest lies in the average treatment effect. Therefore, we suggest including a dummy variable, indicating whether a measurement occasion belongs to a baseline phase (either the first or the second one) or a treatment phase (either the first or the second one). On top of this, a researcher might also be interested in whether the estimated treatment effect is different between AB pairs by including a second dummy coded variable *pair* (indicating the AB pair).

If the research questions focus on the change in outcome score and/or trends between consecutive phases, then we suggest coding three dummy variables indicating the phase (*A1B1*, *B1A2*, and *A2B2*) and four time variables (*Time, Time1, Time2* and *Time3*). We center *Time* around the beginning of the first baseline phase, *Time1* around the beginning of the first treatment phase, *Time2* around the beginning of the second baseline phase and *Time3* around the beginning of the second treatment phase. If a researcher is interested in whether there is a difference between the slope in the first baseline phase and the second baseline phase, we choose to keep *Time* constant during the first treatment phase. As illustrated in the three design matrices, the coding of the time predictors depends on which phase's slopes the researcher wants to compare. A summary table using design matrices 6 through 8 to analyze the ABAB design of the first participant of Moes' study (1998) is given in Table 6.4.

Table 6.4

Summary Results ABAB Regression Analysis using the Alternative Way of Coding: Design Matrix 6 – Design Matrix 8

Coefficient	Parameter	Estimate (*SE*)
Design Matrix 6		
Outcome score at the start of A1	$\hat{\beta}_0$	73.61* (4.49)
Linear trend during A1	$\hat{\beta}_1$	6.55* (1.83)
Immediate treatment effect in the first AB pair	$\hat{\beta}_2$	-51.15* (4.56)
Linear trend during B1	$\hat{\beta}_3$	0.59 (1.83)
Difference in outcome score when removing the treatment	$\hat{\beta}_4$	29.42* (7.56)
Linear trend during A2	$\hat{\beta}_5$	1.52 (1.83)
Immediate treatment effect in the second AB pair	$\hat{\beta}_6$	-53.85* (-6.92)
Difference in trend between B1 and B2	$\hat{\beta}_7$	-1.95 (-0.75)
Design Matrix 7		
Outcome score at the start of A1	$\hat{\beta}_0$	73.61* (4.49)
Linear trend during A1	$\hat{\beta}_1$	6.55* (1.83)
Immediate treatment effect in the first AB pair	$\hat{\beta}_2$	-51.15* (7.56)
Difference in trend between A1 and B1	$\hat{\beta}_3$	-5.96* (2.59)
Difference in outcome score when removing the treatment	$\hat{\beta}_4$	29.42* (7.56)
Linear trend during A2	$\hat{\beta}_5$	1.52 (1.83)
Immediate treatment effect in the second AB pair	$\hat{\beta}_6$	-53.85* (-6.92)
Difference in trend between A2 and B2	$\hat{\beta}_7$	-2.88 (2.59)
Design Matrix 8		
Outcome score at the start of A1	$\hat{\beta}_0$	73.61* (4.49)
Linear trend during A1	$\hat{\beta}_1$	6.55* (1.83)
Immediate treatment effect in the first AB pair	$\hat{\beta}_2$	-51.15* (7.56)
Linear trend during B1	$\hat{\beta}_3$	0.59 (1.83)
Difference in outcome score when removing the treatment	$\hat{\beta}_4$	29.42* (7.56)
Difference in trend between A1 and A2	$\hat{\beta}_5$	-5.03 (2.59)
Immediate treatment effect in the second AB pair	$\hat{\beta}_6$	-53.85* (-6.92)
Difference in trend between B1 and B2	$\hat{\beta}_7$	1.95 (-0.75)

Note. * $p < .05$.

6.4 Alternating Treatment Designs

Thus far, we have only discussed SSEDs in which a single treatment is introduced (e.g., AB design, multiple-baseline designs, and ABAB reversal designs). In many cases however, researchers are not only interested in whether one treatment works but also whether one treatment works better in comparison to another. In an alternating treatment design (ATD), two or more treatments are rapidly alternated (Barlow & Hayes, 1979). In a typical ATD, data collection starts with a baseline phase, but during the treatment phase, two or more treatments are alternated (see Figure 6.13). Because we cannot identify distinct treatment "phases", the analysis of the data of an ATD differs from those for the SSEDs previously discussed. In order to illustrate the design matrices necessary for analyzing ATD data, we use data from the study of Luiselli, Suskin, and McPhee (1981). In this study the authors investigate the effects of an intermittent (i.e. treatment 1; Figure 6.13) versus continuous (treatment 2; Figure 6.13) schedule of overcorrection for a self-injurious autistic child.

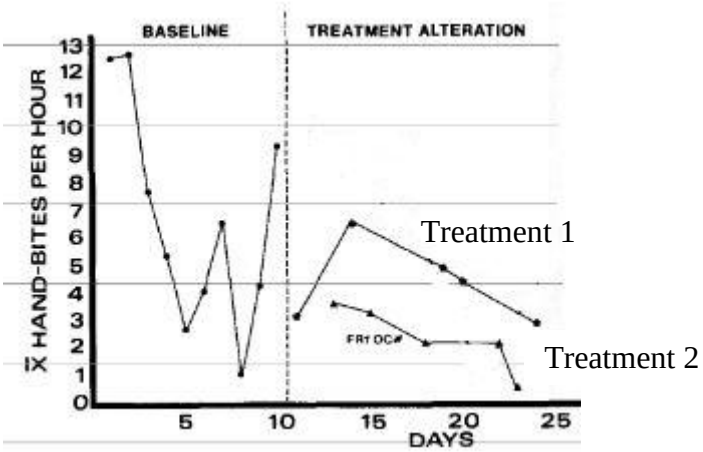

Figure 6.13. Graphical presentation of an alternating treatment design. From "Continuous and intermittent application of overcorrection in a self-injurious autistic child: Alternating treatments design analysis" by Luiselli, J.K., Suskin, L., & McPhee, D.F. (1981). J*ournal of Behavior Therapy and Experimental Psychiatry*, *12*, 355-358.

6.4.1.1 Design matrix 9

Using design matrix 9, an SSED researcher is interested in:

(a) What is the outcome score at the start of the baseline phase and what is the trend during the baseline phase?

(b) What is the immediate treatment effect for treatment 1 and treatment 2, respectively?

(c) What are the changes in slopes between the baseline versus the first treatment and between the baseline versus the second treatment, respectively?

Again we use an extension of the Center et al. (1985-1986) regression approach by introducing dummy variables for each treatment. The dummy coded variables, $Treatment_{mi}s$, indicate the treatment phase ($Treatment_{mi}$ = 1 if the person is in treatment phase *m* on moment *i.*, zero otherwise. If all the $Treatment_{mi}$s are zero, then the measurement occasion belongs to the baseline phase). Extending Equation 6.1 for two treatments, using treatment indicators $Treatment_{1i}$ and $Treatment_{2i}$ results in the following:

$$Y_i = \beta_0 + \beta_1 Time_i + \beta_{21} Treatment_{1i} + \beta_{22} Treatment_{2i} + \beta_{31} Treatment_{1i} Time1_i + \beta_{32} Treatment_{2i} Time2_i + e_i \text{ with } e_i \sim N(0, \sigma_e^2) \quad (6.10)$$

The immediate treatment effect of the first treatment, β_{21}, is the difference between the predicted outcome score using the first treatment's regression model and the predicted outcome score using the baseline regression model at the first measurement occasion of the first treatment. β_{22} is then the immediate treatment effect of the second treatment. Therefore we center the time in $Treatment_{1i} Time1_i$ around the first measurement occasion of the first treatment and $Treatment_{2i} Time2_i$ around the first measurement of the second treatment (see Figure 6.14). Equation 6.10 allows a comparison between the immediate treatment effect β_{21} in the first treatment and the immediate treatment effect in the second treatment β_{22}. Also the treatment effect on the time trends in both treatment phases (i.e. β_{31} and β_{32}) can be tested. The regression coefficients of interest in Equation 6.10 are graphically presented in Figure 6.14. The results using this proposed analysis method using the data from Luiselli et al. (1981) are given in Table 6.5. An interpretation of the coefficients of interest is provided as well.

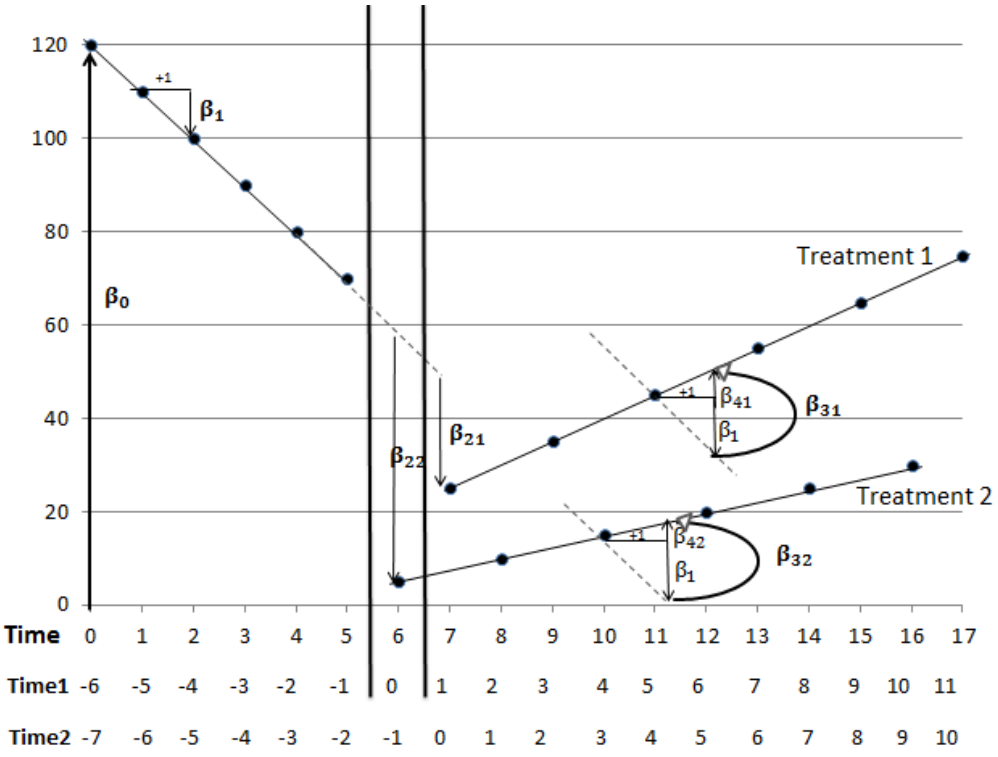

Figure 6.14. Hypothetical alternating treatment design. Graphical presentation of coefficients in design matrix 9. The *X*-axis represents the variables *Time*, *Time1* and *Time2*. *Time* is recoded such that *Time* = 0 for the first measurement occasion in the baseline phase. *Time1* is recoded such that *Time1* = 0 at the start of treatment 1. *Time 2* is recoded such that *Time2* = 0 at the start of treatment 2.

6.4.1.2 Design Matrix 10

The previous design matrix allows estimating the change in slope due to the first treatment, β_{31}, and the change in slope due to the second treatment, β_{32}. In case the research interest lies in estimating the slopes in both treatment phases instead of the changes in slopes, we propose the design matrix in which the baseline phase time variable, *Time*, is held constant during the treatment phases (see Figure 6.15). A graphical presentation of the coefficients using design matrix 10 is given in Figure 6.15. The results using design matrix 10 applied to the data of Luiselli et al. (1981) are displayed in Table 6.5. An interpretation of the coefficients of interest is provided as well.

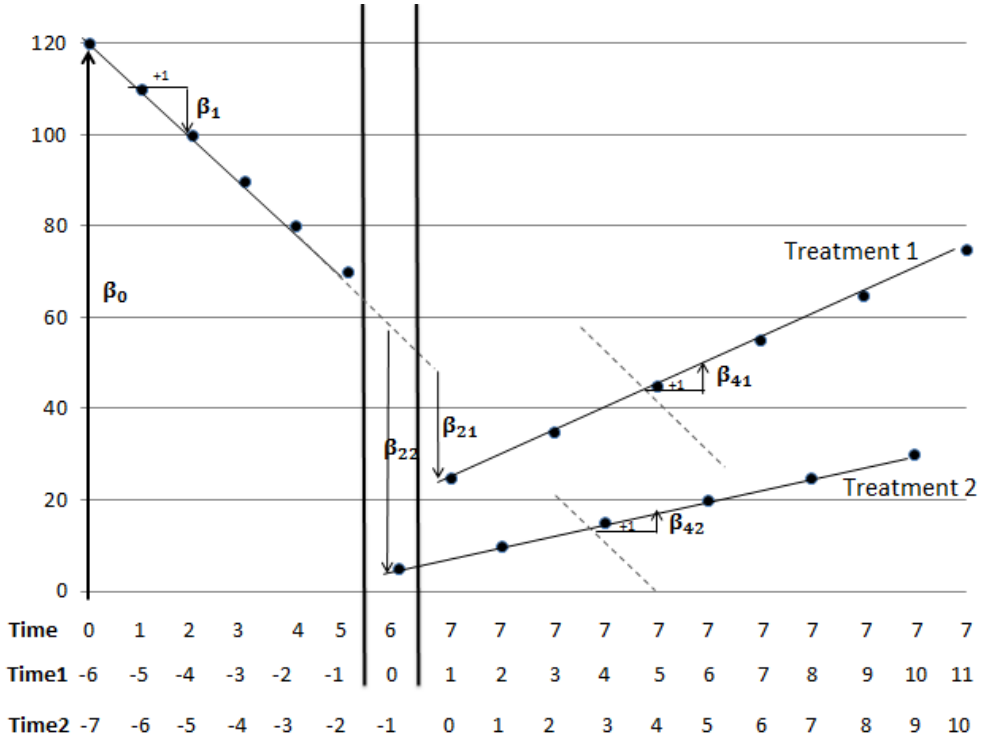

Figure 6.15. Hypothetical alternating treatment design. Graphical presentation of coefficients in design matrix 10. The *X*-axis represents the variables *Time*, *Time1* and *Time2*. *Time* is recoded such that *Time* = 0 for the first measurement occasion in the baseline phase and is kept constant in the treatment phase. *Time1* is recoded such that *Time1* = 0 at the start of treatment 1. *Time2* is recoded such that *Time2* = 0 at the start of treatment 2.

6.4.2 Conclusion alternating treatment designs

Alternating treatment designs are of particular interest if multiple treatments are under investigation. In this design, the treatments are alternatingly given to the participant(s). This allows estimating multiple treatment effects using one design and one participant. In order to analyze this type of data, we suggest to include two dummy coded variables, $Treatment_{1i}$ and $Treatment_{2i}$ in order to estimate the immediate treatment effect for each treatment. Furthermore, three time variables are needed to estimate the linear trend or the change in linear trend. The coding of these time variables depends on the research interest and therefore we suggested two modeling options. The results of design matrix 9 and 10 are summarized in Table 6.5. The only difference between the matrices is the estimate of β_{31} and β_{32} representing the change in slope between treatment and baseline in design matrix 9 versus the slope itself during each treatment phase.

Table 6.5

Summary Results Alternating Treatment Design Regression Analysis: Design Matrix 9 – Design Matrix 10

Coefficient		Estimate (*SE*)
Design Matrix 10		
Outcome score at the start of A1	$\hat{\beta}_0$	7.74* (1.96)
Linear trend during A1	$\hat{\beta}_1$	-0.33 (0.37)
Immediate treatment effect of the first treatment	$\hat{\beta}_{21}$	5.86 (3.62)
Immediate treatment effect of the second treatment	$\hat{\beta}_{22}$	1.42 (3.65)
Difference in trend between the first treatment and the baseline	$\hat{\beta}_{31}$	-1.07 (0.79)
Difference in trend between the second treatment and the baseline	$\hat{\beta}_{32}$	-0.23 (0.68)
Design Matrix 11		
Outcome score at the start of A1	$\hat{\beta}_0$	7.74* (1.96)
Linear trend during A1	$\hat{\beta}_1$	-0.33 (0.37)
Immediate treatment effect of the first treatment	$\hat{\beta}_{21}$	5.86 (3.62)
Immediate treatment effect of the second treatment	$\hat{\beta}_{22}$	1.10 (3.42)
Linear trend during treatment 1	$\hat{\beta}_{41}$	-1.39 (0.70)
Linear trend during treatment 2	$\hat{\beta}_{42}$	-0.56 (0.58)

Note. * $p < .05$.

6.5 Discussion

6.5.1 General conclusion

The main rationale for this article is the result of the growing interest in SSEDs as a means of establishing an evidence base for intervention effects. Due to the increased number of published SSED studies in a variety of different research fields, there is a need to summarize SSED data across studies in order to make generalized decisions and to inform research and policy. If a conclusion is reached across a large number of studies, one can be more confident in the study results. In order to synthesize SSED data across a large number of studies, statistical techniques are needed. In this article we proposed a regression model-based approach, which was already suggested in the 80s, as a flexible and easy way to analyze SSED data. The regression approach results in an effect size estimate (i.e., immediate treatment effect and treatment effect on the slope), and allows the modeling of autocorrelation, non-linear trends, predictors, etc. Despite the enormous flexibility of the regression technique, little is known about the interpretation of the regression coefficients and the consequences of misspecifying the design matrix. As a consequence, this article aims to provide guidance to SSED analysts about how to specify the design matrix in order to answer predefined research question. We illustrated that the specification of the design matrix and the interpretation of the regression coefficients are interdependent. We only discussed the most common and typical research questions per design that SSED researchers might have. Other researcher questions can also be resolved using a little modification of the suggested coding schemes. Also note that parameters of interest can be estimated indirectly. For instance, in design matrix 4, the slope during the treatment is estimated directly but it might be the case that the SSED researcher is interested in the change in slope between baseline and treatment rather than in the slope during the treatment. In this situation, the change in slope can be estimated based on the estimated slope during the baseline and the estimated slope during the treatment. However, we do not advise calculating parameters of interest indirectly, because the standard errors and *p*-values are not given and have to be calculated by hand.

6.5.2 Limitations and suggestions for future research

Although we focused on the design matrix of the three most common SSEDs, we are aware that combinations of different designs are possible. For instance, an alternating treatments design might be implemented using the same staggering of intervention start times across participants. Also simplification or extensions of the presented designs are possible, including, for example, ABA, ABABAB designs or alternating treatment designs without a baseline phase. In this article we present the coding schemes for the most common design types, but these coding schemes can easily be modified. Because we used regression analysis, we have to re-emphasize the assumptions that are made, such as linear trends in baseline and treatment phases, and errors that are normally and independently distributed. A violation of one or some of these assumptions might lead to misleading results. Moreover the regression equation as proposed here is suitable for continuous data, when the data are counts, a generalized (multilevel) regression model would be more appropriate. In the current examples, we included only two kinds of predictors, representing the phase and time, but additional predictors might be included such as a quadratic time term. It is also possible that no time trends are expected. Although we did not cover all the possible variations of design matrices, the ones proposed in this article cover a substantial variety of research questions for the three most common SSEDs. However, our suggestion for further research is to further extend the regression equations proposed in this manuscript by adding complexities such as autocorrelation, non-linear trends, non-continuous outcome scores, heterogeneous within-case variance etc. Further research is needed to combine different types of SSEDs because this can lead to more accurate and reliable treatment effect inferences. If the same research finding is found across different types of SSEDs, the research findings are more reliable. In addition, more SSED data are available, which results in more accuracy in the average treatment effect estimate.

Despite the limitations and assumptions, use of parametric statistics is preferable over nonparametric statistics because the latter approach cannot easily model trends in the data, discriminate between large treatment effects due to ceiling effects, and be associated with a tractable sampling distribution. The regression approach is easily conducted using standard statistical software packages and results in calculation of effect size estimates. The resulting effect size estimates can be used to summarize results from a large body of SSED studies thereby offering a stronger evidence base about interventions' effects which will ultimately be of great use for informing educational research and policy.

Chapter 7|

From a Single-Level to a Multilevel Analysis of Single-Case Experimental Designs[6]

Abstract

Multilevel modeling provides one approach to synthesizing single-case experimental design data. In this study, we present the multilevel model (the two-level and the three-level models) for summarizing single-case results over cases, over studies, or both. In addition to the basic multilevel models, we elaborate on several plausible alternative models. We apply the proposed models to real datasets and investigate to what extent the estimated treatment effect is dependent on the modeling specifications and the underlying assumptions. By considering a range of plausible models and assumptions, researchers can determine the degree to which the effect estimates and conclusions are sensitive to the specific assumptions made. If the same conclusions are reached across a range of plausible assumptions, confidence in the conclusions can be enhanced. We advise researchers not to focus on one model but conduct multiple plausible multilevel analyses and investigate whether the results depend on the modeling options.

Keywords: single-case experimental design, multilevel analysis

[6] This chapter has been published as Moeyaert, M., Ferron, J., Beretvas, S.N.,& Van den Noortgate, W. (2014). From a single level analysis to a multilevel analysis of single-case experimental data. *Journal of School Psychology, 52*, 191-211. doi: http://dx.doi.org/10.1016/j.jsp.2013.11.003

7.1 Introduction

The use of single-case designs in a variety of different research fields in education as well as the suggested methods to analyze these types of designs have been expanding for decades. In this article, we describe and illustrate one method, namely, the use of multilevel modeling, which provides an appropriate method to analyze and summarize single-case data (Moeyaert et al., 2013a; Owens & Ferron, 2012; Van den Noortgate & Onghena, 2008). In a single-case study, usually multiple cases, subjects, or participants are involved and repeatedly measured over time (Shadish & Sullivan, 2011). Therefore, in addition to the case-specific estimates, it is useful to develop methods to summarize the results over cases within a particular study. In the first part of this article, we present the basic two-level regression modeling framework that can be used to estimate the treatment effect across cases within studies and the between-case variance of this treatment effect (Van den Noortgate & Onghena, 2003a). We suggest and illustrate a sensitivity analysis approach in which multiple alternative specifications of this basic two-level model are examined. For illustration, we use the dataset of Lambert, Cartledge, Heward, and Lo (2006). In order to allow further examination of external validity and contribute to evidence-based research (Shadish & Rindskopf, 2007), multiple single-case studies measuring the same outcome variable can be combined using the three-level model, which is a straightforward extension of the two-level model. Thus, the second part of this article focuses on the three-level model. We present the basic three-level model assuming no linear trends in which the treatment effect across cases and across studies can be estimated as well as the between-case and between-study variances of this estimate. We will discuss the flexibility of this three-level modeling framework by suggesting multiple alternatives to the basic three-level model. The basic three-level model and alternative specifications of this basic three-level model will be illustrated by summarizing five studies in which a multiple-baseline across participants design was used to investigate the effects of pivotal response training with children with autism.

7.2 From a Single-Level to a Two-Level Framework

7.2.1 Two-level model

In single-case experiments, usually more than one case is the focus of interest (Shadish & Sullivan, 2011), such as in the replicated ABAB reversal designs and the multiple-baseline across participants designs. In this first design, there are multiple baseline phases (A phases) and multiple treatment phases (B phases), and the same ABAB design is implemented simultaneously to different participants (see Figure 7.1a). In the multiple-baseline across participants design, an AB phase design (with one baseline phase, A, and one treatment phase, B) is delivered simultaneously to different participants and the start of the delivery is staggered across the participants (see Figure 7.1b).

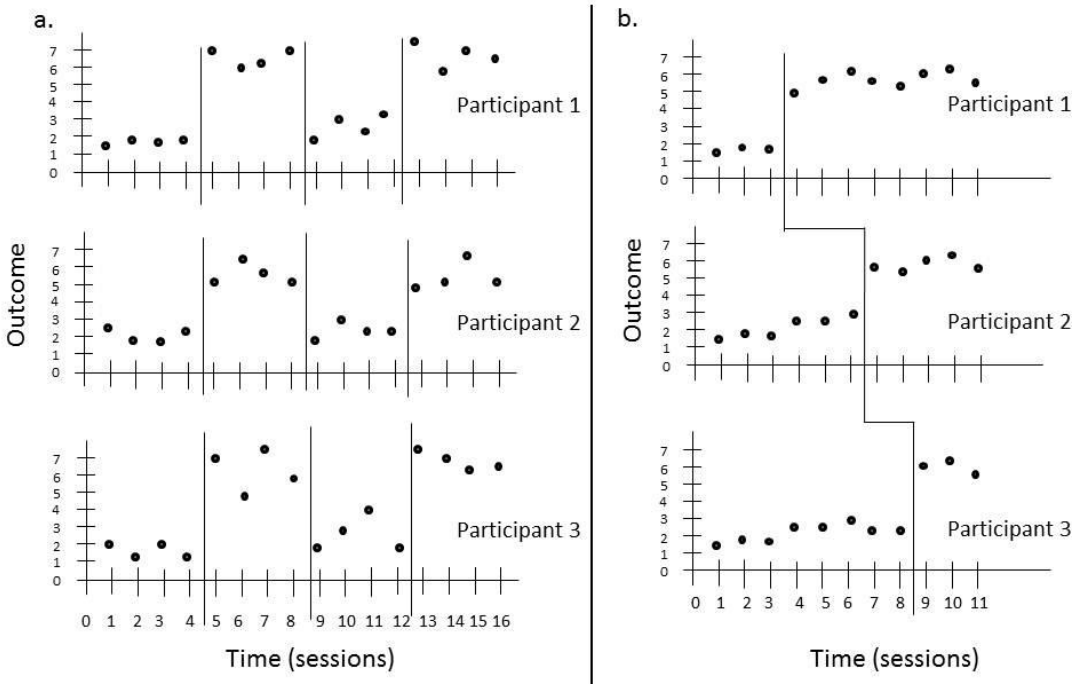

Figure 7.1. Graphical display of an ABAB reversal design (a) and a multiple-baseline across participants design (b) using hypothetical datasets.

In order to analyze these single-case data, an autoregressive integrated moving average approach (Velicer & Fava, 2003), an ordinary least square regression analysis (Huitema & McKean, 1998), or a generalized least squares regression analysis (Maggin et al., 2011) could be performed for each case within the single-case study separately. These analysis procedures allow researchers to estimate case-specific treatment effects. However, in order to add to evidence-based research, researchers are not only interested in whether a specific treatment works for a particular case but also whether its effect can be generalized to other cases. Therefore, in addition to case-specific estimates, there is a need to estimate the average

treatment effect across cases within the same study. If there are only two cases within a study, a single-level analysis is reasonable to estimate the treatment effects for the two cases separately, compare them, and calculate the average in order to find the average treatment effect. However, Shadish and Sullivan's (2011) review of 809 single-case studies published in 2008 indicated that the number of cases within studies can range from 1 to 13 with an average of 3.64. Moreover, there is an increased interest in using scaled-up multiple-baseline designs. For instance, the study of Koutsoftas et al. (2009) included 36 participants, which makes it practically complex and inefficient to estimate treatment effects for each participant separately and to calculate the average treatment estimate and the between- and within-case variability of this treatment effect.

Therefore, Van den Noortgate and Onghena (2003a, 2003b) suggested combining single-case data within a study using a two-level model, which is a simple extension of a regression equation, in which the hierarchical nature of single-case data is taken into account. Measurement occasions, going from 1 up to *I*, are nested within a case, *j*, and in each study there are *J* cases. At the first level, a regression equation in which the outcome score for case *j* at measurement occasion *i*, y_{ij} (e.g., the number of correct responses at a particular moment *i* for case *j*) is regressed on an intercept, indicating the baseline level for case *j* and a dummy coded variable, $Phase_{ij}$, indicating the condition (if $Phase_{ij}$ = 0, measurement occasion *i* belongs to the baseline phase, A, otherwise to the treatment phase, B). Following regression equation can be used:

Level 1 (Model 1A):

$$Y_{ij} = y_{ij} = \beta_{0j} + \beta_{1j}Phase_{ij} + e_{ij} \text{ with } e_{ij} \sim N(0, \sigma_e^2) \tag{7.1}$$

The within-case residuals, the e_{ij}s, are assumed to be independent, identically, and normally distributed. At the second level, the case-specific coefficients from the first level, β_{0j} and β_{1j}, are modeled as varying across participants because it is unlikely that the estimated baseline level and the treatment effect are the same for all cases within a particular study:

Level 2 (Model 1A):

$$\begin{cases} \beta_{0j} = \theta_{00} + u_{0j} \\ \beta_{1j} = \theta_{10} + u_{1j} \end{cases} \text{ with } \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0u_1} \\ \sigma_{u_1u_0} & \sigma_{u_1}^2 \end{bmatrix} \right) \tag{7.2}$$

In Equation 7.2, θ_{00} indicates the average baseline level, and θ_{10} represents the treatment effect across the *J* cases. Each individual case, *j*, can have a baseline level and a

treatment effect that deviate from the average baseline level, θ_{00}, and the average treatment effect, θ_{10}, quantified by the participant-specific residuals (u_{0j} and u_{1j}, respectively). Level-2 residuals are also assumed to be independent and identically, and multivariate normally distributed. Single-case researchers are interested in the average baseline level, θ_{00}, as this level can be used to substantiate the need for intervention. Of primary interest, however, is the average treatment effect, θ_{10}, because this parameter indexes the magnitude of the shift in behavior that tends to occur with intervention. This two-level framework can also be used to estimate the between-case variance in baseline level and treatment effect indicated by $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$ respectively and the covariance between the baseline level and treatment effect, indicated by $\sigma_{u_0u_1}$. The variance component $\sigma^2_{u_1}$ would be particularly useful for a researcher interested in determining whether the shift in behavior associated with treatment is similar across participants or whether the shift in behavior differs substantially across participants and thus indicates that the treatment is differentially effective. Another advantage is that, in addition to estimating the average treatment effect and the variance in the treatment effect, researchers can obtain empirical Bayes estimates of case-specific treatment effects.

By using this two-level model, we have to be aware of several assumptions. First, we assume that the outcome variable is continuous (e.g., the score on a math test) and that the errors at the different levels are independent, identically, and normally distributed. Another drawback is that the variance estimates (i.e., the between-case variance of the baseline level and the between-case variance of the treatment effect) can be biased when a limited number of participants are included. Ferron et al. (2009) studied restricted maximum likelihood estimation of this two-level model assuming no covariance between the baseline level and treatment effect (i.e., $\sigma_{u_0u_1} = 0$) and found unbiased estimates of the average treatment effect but biases in the estimates of $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$ with four, six, and eight participants. Furthermore, the model does not take trends into account whereas linear, quadratic, or nonlinear trends are possible. Modeling a time trend can be accomplished by adding predictors at the first level, for instance a continuous time variable if a linear trend is expected. Also, case-specific predictors, such as age or gender, can be included at the second level in order to explain the between-case variability. We illustrate the flexibility of the two-level model by proposing several modeling options in addition to the basic two-level model (see Equation 7.1). The basic two-level model together with several alternative models will be illustrated using the Lambert et al. dataset (2006). By analyzing this dataset using different models, we can also investigate to what extent the estimated treatment effect, which is the primary interest of the

single-case researcher, is sensitive to the different modeling options. If we will find similar results across the different models, then we can be more confident in the results.

7.2.2 Empirical illustration of the two-level model

As discussed in the first section, the multilevel modeling approach is very flexible which gives us several modeling options for analyzing the Lambert et al. (2006) dataset. In the first part, we illustrate the basic two-level model, which will be modified in several ways in the second part, representing more complex and probably more realistic modeling assumptions. When discussing the results, we use .05 as the alpha-level. We used SAS 9.3 to conduct the analysis and the SAS codes for the basic two-level model (i.e., Model 1) as well as the extensions to this model (Models 2 to 4) contained in Addendum A5.

7.2.2.1 Model 1: the basic two-level model

We use the replicated ABAB reversal design study of Lambert et al. (2006) to illustrate the two-level model. We indicate the first and the second baseline phases by A1 and A2, respectively, and the first and second treatment phases by B1 and B2, respectively (see Figure 7.2).

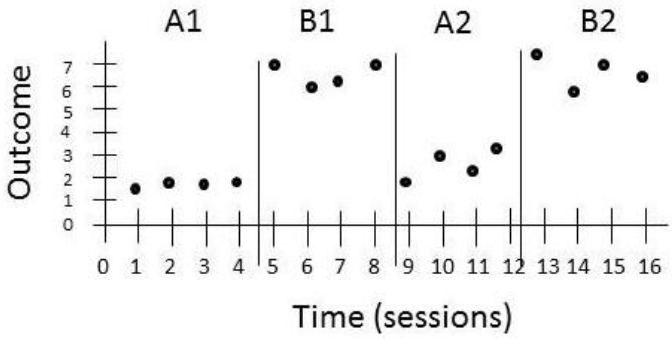

Figure 7.2. Graphical display of an ABAB reversal design using a hypothetical dataset.

In the simplest scenario, the single-case researchers' interest lies in the average estimated treatment effect across cases within a study and the variability of this estimated effect between cases. In this scenario, measurement occasions belonging to baseline phases (A1 or A2), are indicated by $Phase_{ij}$ = 0, and measurements obtained during treatment phases (B1 or B2) have $Phase_{ij}$ = 1. The average estimated baseline level, $\hat{\theta}_{00}$, the average estimated treatment effect, $\hat{\theta}_{10}$, the between-case variance of these estimates and the covariance between these estimates as well as the within-case variance estimate are presented in Table 7.1 and labeled as Model 1A. The SAS code can be found in Addendum A5 (Model 1A).

Table 7.1

Parameter and Standard Error Estimates Resulting from Estimation of Model 1A and Model 1B Using the Lambert et al. (2006) Dataset

		Parameter	Parameter estimate	*SE*	*p*
Model 1A		Fixed coefficient			
	Average baseline level	θ_{00}	6.78*	0.40	< .001
	Average treatment effect	θ_{10}	-5.40*	0.34	< .001
		(Co)variance component			
	Baseline level	$\sigma^2_{u_0}$	1.14*	0.67	.045
	Treatment effect	$\sigma^2_{u_1}$	0.43	0.49	.191
	Covariance between baseline level and treatment effect	$\sigma_{u_0u_1}$	-0.79	0.54	.142
	Residual variance	σ^2_e	4.44*	0.40	< .001
Model 1B		Fixed coefficient			
	Average baseline level, first AB pair	θ_{00}	6.88*	0.34	< .001
	Average treatment effect, first AB pair	θ_{10}	-5.66*	0.38	< .001
	Average change in level, from B1 to A2	θ_{20}	5.35*	0.61	< .001
	Average treatment effect, second AB pair	θ_{30}	-5.08*	0.49	< .001
		Variance component			
	Baseline level, first AB pair	$\sigma^2_{u_0}$	0.52	0.41	.102
	Treatment effect, first AB pair	$\sigma^2_{u_1}$	0.00	-	-
	Change in level, from B1 to A2	$\sigma^2_{u_2}$	1.93	1.51	.100
	Treatment effect, second AB pair	$\sigma^2_{u_3}$	1.07	1.02	.148
	Residual variance	σ^2_e	4.28*	0.39	< .001

Note. * $p < .05$.

In a second scenario, the single-case researcher is interested in the estimated change in outcome score when another phase is introduced. In order to estimate the change in outcome score due to the introduction or removal of a treatment, Shadish et al. (2013) suggested extending Equation 7.1 by adding three dummy coded predictors indicating the phase. We chose to name the dummy variables $A1B1_{ij}$, $B1A2_{ij}$ and $A2B2_{ij}$. The first dummy variable $A1B1_{ij}$ equals 1 if measurement occasion i from case j is obtained after the first baseline phase; $B1A2_{ij}$ equals 1 for all measurement occasions after the first treatment phase, and $A2B2_{ij}$ equals 1 if the measurement occasion occurs in the last treatment phase. If $A1B1_{ij}$, $B1A2_{ij}$, and $A2B2_{ij}$ equal simultaneously 0, then the measurement is taken in the first baseline phase. By choosing this way of coding, the expected value during the first baseline phase (A1) equals β_{0j} (i.e., $\beta_{0j} + \beta_{1j}* 0 + \beta_{2j}* 0 + \beta_{3j} * 0$), whereas the expected value during the first treatment phase equals $\beta_{0j} + \beta_{1j}$ (i.e., $\beta_{0j} + \beta_{1j}* 1 + \beta_{2j}* 0 + \beta_{3j} * 0$). Therefore, β_{1j} indicates the treatment effect during the first treatment phase. The expected value during the second baseline phase is $\beta_{0j} + \beta_{1j} + \beta_{2j}$ (i.e., $\beta_{0j} + \beta_{1j}* 1 + \beta_{2j}* 1 + \beta_{3j} * 0$), and in this way, β_{2j} indicates the effect of removing the treatment on the outcome score.

The expected value during the second intervention is $\beta_{0j} + \beta_{1j} + \beta_{2j} + \beta_{3j}$ (= $\beta_{0j} + \beta_{1j}* 1 + \beta_{2j}* 1 + \beta_{3j} * 1$), and therefore β_{3j} is the treatment effect during the second AB pair.
This results in Equation 7.3:

Level 1 (Model 1B):

$$Y_{ij} = \beta_{0j} + \beta_{1j}A1B1_{ij} + \beta_{2j}B1A2_{ij} + \beta_{3j}A2B2_{ij} + e_{ij} \text{ with } e_{ij} \sim N(0, \sigma_e^2) \quad (7.3)$$

In Table 7.2, the coding scheme for the first case of the Lambert et al. (2006) study is demonstrated. Using these three dummy variables (i.e., $A1B1_i$, $B1A2_i$, and $A2B2_i$), β_{1j} indicates the change in level between phase A1 and B1, β_{2j} refers to the jump from phase B1 to phase A2, and the last coefficient, β_{3j}, represents the change in expected outcome score from phase A2 to phase B2. The four coefficients of the first level vary at the second level, which makes it possible to estimate the average treatment effects across cases and the between-case variability in this treatment effect. The results of using this second way to analyze the single-case data are presented in Table 7.1 under Model 1B and the SAS code can be found in Addendum A5 (Model 1B). Note that for both Model 1A and Model 1B the covariance between the coefficients at the second level is estimated. However, we only presented the covariance for Model 1 A because otherwise Table 7.1 would be too extensive, and the main interest lies in the fixed effects and the variance estimates and not in the covariance estimates.

Table 7.2

Demonstrating Second Way of Coding Predictors in an ABAB Reversal Design using Model 1B

A1B1	B1A2	A2B2	Y
0	0	0	7
0	0	0	9
0	0	0	8
0	0	0	6
0	0	0	7
0	0	0	4
0	0	0	5
0	0	0	1
1	0	0	2
1	0	0	0
1	0	0	1
1	0	0	0
1	0	0	0
1	1	0	3
1	1	0	8
1	1	0	8
1	1	0	6
1	1	0	10
1	1	0	10
1	1	0	10
1	1	0	8
1	1	1	3
1	1	1	4
1	1	1	1
1	1	1	3
1	1	1	2
1	1	1	4
1	1	1	0
1	1	1	1
1	1	1	0

For Model 1A, the estimated average treatment effect across phases and across cases was -5.40, $t(16.99) = -15.96$, $p < .001$, indicating a significant reduction in disruptive behavior due to the treatment. From Model 1B, the change in level during the first AB pair and the second AB pair are both statistically significant: $\hat{\theta}_{10} = -5.66$, $t(239) = -14.75$, $p < .001$, and $\hat{\theta}_{20} = -5.08$, $t(8.91) = -10.37$, $p < .001$. The mean of the estimated treatment effects of the first AB pair and the second AB pair is -5.37: [i.e., -5.66 + (-5.08)] / 2 and equals (as expected) approximately the average estimated treatment effect across phases and across cases using Model 1A ($\hat{\theta}_{10} = -5.40$).

In terms of the variance estimates, only the residual within-case variance is statistically significant in both Models. Note that the Wald test was used to investigate whether the variance components were significant. Given the small number of participants, the Wald test is questionable, and it might be better to consider the likelihood ratio test (Snijders & Bosker,

2002). Therefore, the difference in deviance score between the model with the variance component of interest and the model without the variance component of interest can be calculated. For instance, the deviance score of Model 1A without the between-case variance of the baseline level and with the between-case variance of the baseline level equals 1175.2 and 1153.4 respectively. The difference in the deviance is 21.8, which can be compared to a χ^2 distribution with 2 degrees of freedom (i.e., the number of degrees of freedom is calculated as the difference in parameters in the models that are compared) and indicates a statistically significant between-case variance of the intercept (similar to what was found with the Wald test, see Table 7.1). In the remainder of this article, we will focus primarily on the average effects, but we will present the variance components for completeness and use the Wald test for simplicity (because it is used by default in the statistical software program we used). We encourage readers to view the estimates of the variance components and the inferences about them with more caution than the estimates of the average effects and the inferences about them.

Although we estimated the average baseline level and average treatment effect across cases, we can also estimate the case-specific baseline level and treatment effect. Therefore, we simply add the command "solution" after the random statement in the model specification (see Addendum A5, Model 1A and Model 1B). Table 7.3 presents the results for the first three cases of the Lambert et al. (2006) dataset for Models 1A and 1B, respectively. Using Model 1A, $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ refer to the estimated baseline level and the treatment effect respectively for the j^{th} case. Using Model 1B, $\hat{\beta}_{0j}$ refers to the estimated baseline level during the first baseline for the j^{th} case and $\hat{\beta}_{1j}$, $\hat{\beta}_{2j}$, and $\hat{\beta}_{3j}$ refer to the changes in level between the consecutive phases for the j^{th} case. The estimates using Model 1B are graphically presented in Figure 7.3.

Table 7.3

Results Empirical Bayes Estimation of the Case-Specific Effects for the First Three Cases of the Lambert et al. (2006) Dataset using the Basic Two-Level Model

		Parameter	Parameter estimate	*SE*	*p*
Model 1A					
Case 1	Baseline level	β_{01}	6.79*	0.52	< .001
	Treatment level	β_{11}	-5.23*	0.40	< .001
Case 2	Baseline level	β_{02}	8.09*	0.54	< .001
	Treatment level	β_{12}	-6.29*	0.79	< .001
Case 3	Baseline level	β_{03}	7.80*	0.58	< .001
	Treatment level	β_{13}	-5.81*	0.81	< .001
Model 1B					
Case 1	Baseline level A1	β_{01}	5.73*	0.72	< .001
	Treatment level B1	β_{11}	-4.79*	1.14	< .001
	Baseline level A2	β_{21}	6.69*	1.14	< .001
	Treatment level B2	β_{31}	-5.52*	0.98	< .001
Case 2	Baseline level A1	β_{02}	7.22*	0.77	< .001
	Treatment level B1	β_{12}	-5.58*	1.12	< .001
	Baseline level A2	β_{22}	6.96*	1.08	< .001
	Treatment level B2	β_{32}	-6.48*	0.98	< .001
Case 3	Baseline level A1	β_{03}	7.56*	0.83	< .001
	Treatment level B1	β_{13}	-6.76*	1.21	< .001
	Baseline level A2	β_{23}	6.93*	1.17	< .001
	Treatment level B2	β_{33}	-4.76*	1.04	< .001

Note. * $p < .05$.

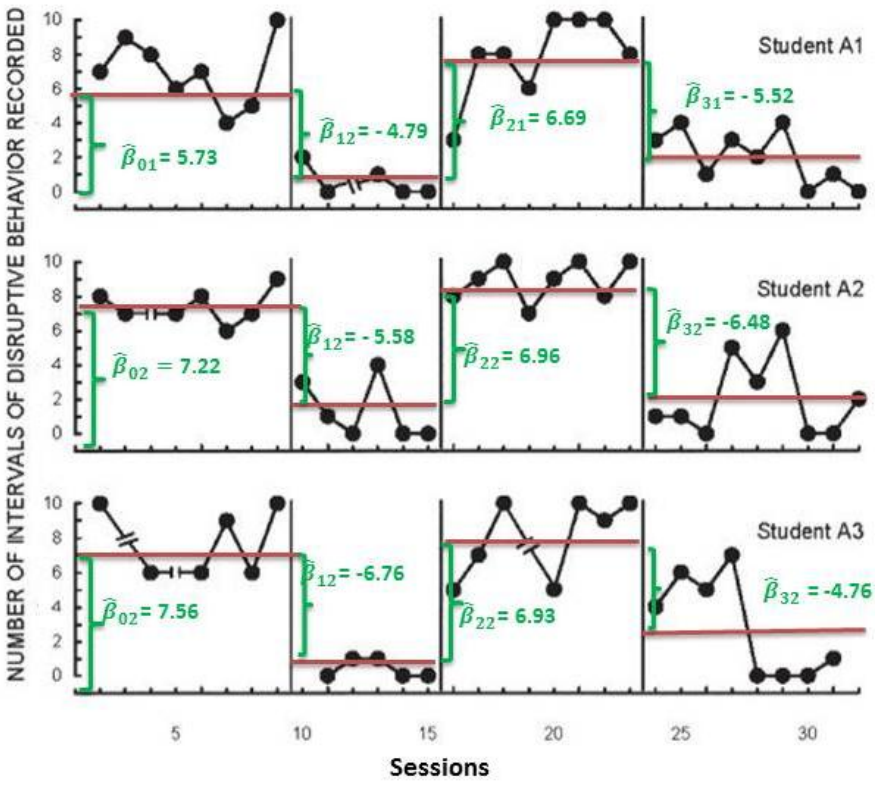

Figure 7.3. Graphical presentation case specific baseline level and changes in level between consecutive phases for the first three cases from the Lambert et al. (2006) study.

Using Models 1A and 1B, we make several assumptions, such as (1) the outcome variable, Y_{ij}, is continuous, (2) errors at the first and second level are independent, identically, and multivariate normally distributed, (3) there are no time-trends, and (4) there is no systematic variation between the two classes from which the participants came. Because we make several assumptions in the basic two-level model, we suggest multiple alternatives to analyze the Lambert et al. (2006) dataset based on visual analysis of the graphs included in the original study. Similar to the first part, in the alternative models, Model A will refer to the first way of coding in which the research interest lies in the average treatment effect estimate across phases and Model B will refer to the alternative way of coding in which the treatment effect during the first AB pair and the second AB pair are estimated separately.

In all these models, we discuss the average estimates across cases for the fixed effects. Case-specific estimates can also be obtained by adding the command "solution" in the random specification (see Addendum A5). In addition to the fixed effect estimates (i.e., treatment effect estimate), the between-case variance in intercept and treatment effects is estimated. Also, the covariances between the regression coefficients at the second level are estimated but not presented in the tables for simplicity.

7.2.2.2 Model 2

In a single-case design, cases are measured repeatedly over time. Therefore, it is likely that outcome scores that are measured closer in time are more related to each other than outcome scores measured further away in time. For instance, in single-case data, event effects that influence the score at a certain moment can also influence scores on one or more succeeding occasions which lead to similarity among errors that are close to each other in time (Kromrey & Foster-Johnson, 1996). As a consequence, the assumption of independence of errors may be violated because of autocorrelation (Ferron et al., 2009; Huitema & McKean, 1994; McKnight, McKean, & Huitema, 2000). In the basic two-level model (see Equation 7.1), we modeled the level-one errors as $\sigma^2 I$, but there are many other covariance structures possible, of which the first-order autoregressive type is often used (Goldstein, 1995; Goldstein, Healey, & Rasbash, 1994; Wolfinger, 1996).

In addition to modeling autocorrelation, we also question the assumption of homogeneous within-case variance across phases. From the graphical presentation of the single-case data (Lambert et al., 2006), we expect that there is more variability in outcome scores during the baseline phase in comparison to the outcome scores during the treatment phase. Therefore, in this model, we assume heterogeneous phase variances and this is indicated by $\sigma^2_{e(A)}$ and $\sigma^2_{e(B)}$ in Table 7.4 referring to the within-case variance in the baseline and treatment phases, respectively. Assuming first-order autoregressive autocorrelation and heterogeneous within-case variance, we obtain the results presented in Table 7.4. $\rho_{(A)}$ and $\rho_{(B)}$ refer to the estimated autocorrelation in the baseline and treatment phases, respectively. The SAS code for Models 2A and 2B is presented in Addendum A5.

Table 7.4

Parameter and Standard Error Estimates Resulting from Estimation of Models 2A and 2B Using the Lambert et al. (2006) Dataset

	Parameter	Parameter estimate	*SE*	*p*
Model 2A	Fixed coefficient			
Average baseline level	θ_{00}	6.73*	0.43	< .001
Average treatment effect	θ_{10}	-5.36*	0.37	< .001
	Variance component			
Baseline level	$\sigma^2_{u_0}$	0.95	0.84	.130
Treatment effect	$\sigma^2_{u_1}$	0.20	0.64	.378
Residual variance, baseline	$\sigma^2_{e(A)}$	6.09*	0.91	< .001
Residual variance, treatment	$\sigma^2_{e(B)}$	3.14*	0.45	< .001
Autocorrelation, baseline	$\rho_{(A)}$	0.34*	0.10	< .001
Autocorrelation, treatment	$\rho_{(B)}$	0.23*	0.10	.019
Model 2B	Fixed coefficient			
Average baseline level, first AB pair	θ_{00}	6.95*	0.42	< .001
Average treatment effect, first AB pair	θ_{10}	-5.76*	0.50	< .001
Average change in level, from B1 to A2	θ_{20}	5.23*	0.55	< .001
Average treatment effect, second AB pair	θ_{30}	-4.92*	0.49	< .001
	Variance component			
Baseline level, first AB pair	$\sigma^2_{u_0}$	0.21	0.62	.362
Treatment effect, first AB pair	$\sigma^2_{u_1}$	0.00	-	-
Change in level, from B1 to A2	$\sigma^2_{u_2}$	1.42	1.85	.221
Treatment effect, second AB pair	$\sigma^2_{u_3}$	0.13	1.16	.455
Residual variance, baseline	$\sigma^2_{e(A)}$	5.96*	0.93	< .001
Residual variance, treatment	$\sigma^2_{e(B)}$	3.22*	0.52	< .001
Autocorrelation, baseline	$\rho_{(A)}$	0.32*	0.10	.002
Autocorrelation, treatment	$\rho_{(B)}$	0.25*	0.11	.027

Note. * $p < .05$

The estimated variance in the treatment phase is smaller (more than twice) than the variance estimated for the baseline phase and both variance estimates are statistically significant. Furthermore, we found that the estimated autocorrelation in the baseline phase and the treatment phase is similar across the two models. The autocorrelation in the baseline and in the treatment phases using Model 2B equals 0.32, $Z = 3.10$, $p = .002$, and 0.25, $Z = 2.21$, $p = .027$, respectively. The estimated treatment effects are similar to those estimated in previous models, which indicates that for this dataset, specification of autocorrelation and across-phase heterogeneity does not have a large influence.

7.2.2.3 Model 3

The graphical presentation of the data of the students investigated in the Lambert et al. (2006) dataset indicates that linear trends during both baseline and treatment phase are possible. Therefore, we suggest including time predictors in Model B in order to investigate changes in slopes due to the transition from one phase to another phase. Moeyaert et al. (2014b) proposed including four time variables (*T1, T2, T3*, and *T4*) in addition to the dummy variables (A1B1, B1A2, and A2B2) to estimate changes in trends between the phases of interest. In this way, single-case researchers can (in addition to modeling a shift in level due to the introduction or removal of a treatment) investigate whether there is a difference in trends between pairs of adjacent phases. The coding for the time variables depends on the changes in trends a researcher is interested in. In this third proposed model, we discuss the coding scheme to investigate whether the treatment effect on the trend during the first AB pair is different than the treatment effect on the time trend during the second AB pair. Other coding schemes are also possible. For a detailed discussion of these alternative coding scenarios, we refer readers to Moeyaert et al. (2014b).

We code *T1, T2, T3* and *T4* as follows: The first time variable, *T1*, equals zero at the start of the first baseline phase (A1) and remains constant after condition B1. *T*2 is centered around the start of the first treatment phase (B1) and remains constant after that phase is completed. *T*3 is centered around the first measurement occasion of the second baseline phase (A2), and *T*4 is centered around the first measurement occasion of the second treatment phase (B2). In Table 7.5, the coding scheme is displayed using one student from the study of Lambert et al. (2006).

Table 7.5

Coding Scheme for the Reversal ABAB Single-Case Design Including Changes in Trends

Dummy coded variable identifying the condition			Time variable				Outcome score
A1B1	B1A2	A2B2	T1	T2	T3	T4	Y
0	0	0	0	-10	-16	-23	10
0	0	0	1	-9	-15	-22	6
0	0	0	2	-8	-14	-21	9
0	0	0	3	-7	-13	-20	4
0	0	0	4	-6	-12	-19	5
0	0	0	5	-5	-11	-18	9
0	0	0	6	-4	-10	-17	6
0	0	0	7	-3	-9	-16	10
0	0	0	8	-2	-8	-15	9
0	0	0	9	-1	-7	-14	9
1	0	0	10	0	-6	-13	4
1	0	0	11	1	-5	-12	3
1	0	0	12	2	-4	-11	4
1	0	0	13	3	-3	-10	4
1	0	0	14	4	-2	-9	1
1	0	0	15	5	-1	-8	0
1	1	0	15	6	0	-7	3
1	1	0	15	6	1	-6	5
1	1	0	15	6	2	-5	8
1	1	0	15	6	3	-4	10
1	1	0	15	6	4	-3	10
1	1	0	15	6	5	-2	10
1	1	0	15	6	6	-1	6
1	1	1	15	6	7	0	3
1	1	1	15	6	8	1	0
1	1	1	15	6	9	2	2
1	1	1	15	6	10	3	4
1	1	1	15	6	11	4	1
1	1	1	15	6	12	5	0
1	1	1	15	6	13	6	1
1	1	1	15	6	14	7	3
1	1	1	15	6	15	8	0
1	1	1	15	6	16	9	1
1	1	1	15	6	17	10	0

In order to estimate the parameters of interest, the following regression equation can be used:

Level 1 (Model 3):

$$Y_{ij} = (\beta_{0j} + \beta_{1j}T1_{ij}) + (\beta_{2j} + \beta_{3j}T2_{ij})A1B1_{ij} + (\beta_{4j} + \beta_{5j}T3_{ij})B1A2_{ij} + (\beta_{6j} + \beta_{7j}T4_{ij})A2B2_{ij} + e_{ij(m)} \text{ with } e_{ij(m)} \sim N(0, \sigma_e^2) \quad (7.4)$$

The *m* in the error term, $e_{ij(m)}$, equals A or B and is used to indicate that we model heterogeneous within-case phase variances. $e_{ij(A)}$ is the residual within the baseline phase and $e_{ij(B)}$ indicates the residual in the treatment phase. β_{0j} and β_{1j} indicate the outcome score at the start of phase A1 and the trend during phase A1, respectively. β_{2j} and β_{3j} represent the immediate treatment effect (i.e., the shift in level at the time of the first treatment phase observation) and the treatment effect on the trend (i.e., the change in slope) in the first AB pair. β_{4j} is the difference in outcome score when removing the treatment (from phase B1 to phase A2), and β_{5j} is the trend during phase A2. β_{6j} is the immediate treatment effect in the second AB pair and β_{7j} is the difference in trend between phase A2 and phase B2. The single-case researcher is especially interested in β_{2j}, β_{3j}, β_{6j} and β_{7j} because β_{2j} and β_{3j} represent the immediate treatment effect and the treatment effect on the slope, respectively, during the first AB pair. β_{6j} and β_{7j} represent the immediate treatment effect and the treatment effect on the slope, respectively, during the second AB pair.

We only discuss the second way of coding (Model B) because it is reasonable that the slopes during similar phases differ. The SAS code is presented in Addendum A5, and the results are displayed in Table 7.6. The covariance between the coefficients is estimated, but for simplicity not included in Table 7.6. Similar to Models 1 and 2, we conclude that the estimated immediate treatment effect during both AB pairs is statistically significant. The treatment effect on the trend during both AB pairs is statistically significant. The estimated autocorrelation during the baseline phase is statistically significant and equals 0.37, $Z = 4.05$, $p < .001$, whereas the estimated autocorrelation during the treatment phase is not significant and equals 0.10, $Z = 1.10$, $p = .269$.

Table 7.6

Parameter and Standard Error Estimates Resulting from Estimation of Model 3 Using the Lambert et al. (2006) Dataset

		Parameter	Parameter estimate	*SE*	*p*
Model 3		Fixed coefficient			
	Average baseline level, phase A1	θ_{00}	7.11*	0.48	< .001
	Average trend, phase A1	θ_{10}	-0.04	0.07	.546
	Average treatment effect, first AB pair	θ_{20}	-5.10*	0.69	<.001
	Average treatment effect on trend, first AB pair	θ_{30}	-0.16*	0.07	.032
	Average change in level from B1 to A2	θ_{40}	4.61*	0.78	< .001
	Average trend, phase A2	θ_{50}	0.49*	0.17	.005
	Average treatment effect, second AB pair	θ_{60}	-5.77*	0.84	< .001
	Average treatment effect on trend, second AB pair	θ_{70}	-0.66*	0.17	<.001
		Variance component			
	Baseline level, phase A1	$\sigma^2_{u_0}$	0.03	0.10	0.40
	Trend, phase A1	$\sigma^2_{u_1}$	0	-	-
	Treatment effect, first AB pair	$\sigma^2_{u_2}$	0	-	-
	Treatment effect on trend, first AB pair	$\sigma^2_{u_3}$	0	-	-
	Change in level from B1 to A2	$\sigma^2_{u_4}$	0	-	-
	Trend, phase A2	σ^2_5	0	-	-
	Treatment effect, second AB pair	$\sigma^2_{u_6}$	0	-	-
	Treatment effect on trend, second AB pair	$\sigma^2_{u_7}$	0	-	-
	Residual variance, baseline	$\sigma^2_{e(A)}$	6.27*	0.92	< .001
	Residual variance, treatment	$\sigma^2_{e(B)}$	2.66*	0.34	< .001
	Autocorrelation, baseline	$\rho_{(A)}$	0.37*	0.09	< .001
	Autocorrelation, treatment	$\rho_{(B)}$	0.10	0.09	.269

Note. * $p < .05$.

7.2.2.4 Model 4

In all previous suggested models, only level-1 predictors were included (i.e., dummy variables indicating the phase to which a measurement occasion belongs and time-related predictors). In this fourth model, we add a predictor at the second level, namely the class to which a participant belongs. The intent of adding a predictor at the second level is to explain the between-case variance. We include the predictor *class* in the first equation of the level-2 equations (see Equation 7.5) in order to explain between-case variance in baseline levels. We expect that the initial outcome score (i.e., outcome score during phase A1) can partially be explained by the class to which a student belongs. We do not expect that changes in outcome scores due to the introduction or removal of the treatment can be explained by the *class* predictor. As a consequence, the regression equations at the second level look as follow:

Level 2 (Model 4):

$$\begin{aligned}\beta_{0j} &= \theta_{00} + \theta_{01}(class)_j + u_{0j}\\ \beta_{1j} &= \theta_{10} + u_{1j}\\ \beta_{2j} &= \theta_{20} + u_{2j}\\ \beta_{3j} &= \theta_{30} + u_{3j}\\ \beta_{4j} &= \theta_{40} + u_{4j}\\ \beta_{5j} &= \theta_{50} + u_{5j}\\ \beta_{6j} &= \theta_{60} + u_{6j}\\ \beta_{7j} &= \theta_{70} + u_{7j}\end{aligned} \quad \text{with} \begin{bmatrix}u_{0j}\\ u_{1j}\\ u_{2j}\\ u_{3j}\\ u_{4j}\\ u_{5j}\\ u_{6j}\\ u_{7j}\\ u_{8j}\end{bmatrix} N\sim \left(\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0u_1} & \sigma_{u_0u_2} & \sigma_{u_0u_3} & \sigma_{u_0u_4} & \sigma_{u_0u_5} & \sigma_{u_0u_6} & \sigma_{u_0u_7} \\ \sigma_{u_1u_0} & \sigma^2_{u_1} & \sigma_{u_1u_2} & \sigma_{u_1u_3} & \sigma_{u_1u_4} & \sigma_{u_1u_5} & \sigma_{u_1u_6} & \sigma_{u_1u_7} \\ \sigma_{u_2u_0} & \sigma_{u_2u_1} & \sigma^2_{u_2} & \sigma_{u_2u_3} & \sigma_{u_2u_4} & \sigma_{u_2u_5} & \sigma_{u_2u_6} & \sigma_{u_2u_7} \\ \sigma_{u_3u_0} & \sigma_{u_3u_1} & \sigma_{u_3u_2} & \sigma^2_{u_3} & \sigma_{u_3u_4} & \sigma_{u_3u_5} & \sigma_{u_3u_6} & \sigma_{u_3u_7} \\ \sigma_{u_4u_0} & \sigma_{u_4u_1} & \sigma_{u_4u_2} & \sigma_{u_4u_3} & \sigma^2_{u_4} & \sigma_{u_4u_5} & \sigma_{u_4u_6} & \sigma_{u_4u_7} \\ \sigma_{u_5u_0} & \sigma_{u_5u_1} & \sigma_{u_5u_2} & \sigma_{u_5u_3} & \sigma_{u_5u_4} & \sigma^2_{u_5} & \sigma_{u_5u_6} & \sigma_{u_5u_7} \\ \sigma_{u_6u_0} & \sigma_{u_6u_1} & \sigma_{u_6u_2} & \sigma_{u_6u_3} & \sigma_{u_6u_4} & \sigma_{u_6u_5} & \sigma^2_{u_6} & \sigma_{u_6u_7} \\ \sigma_{u_7u_0} & \sigma_{u_7u_1} & \sigma_{u_7u_2} & \sigma_{u_7u_3} & \sigma_{u_7u_4} & \sigma_{u_7u_5} & \sigma_{u_7u_6} & \sigma^2_{u_7} \end{bmatrix} \right) \tag{7.5}$$

We only discuss Model B because it is reasonable that the slopes during similar phases differ, and therefore pooling the data from different phases together is conceptually not reasonable. The results of this fourth model are presented in Table 7.7. The covariance between the coefficients is estimated but for simplicity not presented in Table 7.7. The SAS code is presented in Addendum A5, Model 4 .

The baseline level for students belonging to class A equals 7.22, $t(47.1) = 15.24$, $p < .001$, whereas the baseline level for students belonging to class B was estimated to be lower but not by a statistically significant amount: $\hat{\theta}_{01}$ = -0.52, $t(20.1) = -1.45$, $p = .163$. This is consistent with the graphical presentation of the single-case data of the study of Lambert et al. (2006). Note that the treatment effect estimates (immediate shifts in level and changes in slope) and resulting conclusions are similar whether the predictor *class* is added to the model (Table 7.7) or not (Table 7.6).

Table 7.7

Parameter and Standard Error Estimates Resulting from Estimation of Model 4 Using the Lambert et al. (2006) Data

		Parameter	Parameter estimate	*SE*	*p*
Model 4		Fixed coefficient			
	Average baseline level, first AB pair	θ_{00}	7.22*	0.47	< .001
	Average effect of belonging to *class* B during A1	θ_{01}	-0.52	0.36	.163
	Average trend, phase A1	θ_{10}	0.01	0.08	.907
	Average treatment effect, first AB pair	θ_{20}	-5.53*	0.74	< .001
	Average treatment effect on trend, first AB pair	θ_{30}	-0.13	0.07	.085
	Average change in level, from B1 to A2	θ_{40}	4.49*	0.77	< .001
	Average trend, phase A2	θ_{50}	0.39*	0.18	.037
	Average treatment effect, second AB pair	θ_{60}	-5.60*	0.84	< .001
	Average treatment effect on trend, second AB pair	θ_{70}	-0.60*	0.18	<.001
		Variance component			
	Baseline level, first AB pair	$\sigma^2_{u_0}$	0.006	0.09	.472
	Trend, phase A1	$\sigma^2_{u_1}$	0.00	-	-
	Treatment effect, first AB pair	$\sigma^2_{u_2}$	0.00	-	-
	Treatment effect on trend, first AB pair	$\sigma^2_{u_3}$	0.00	-	-
	Change in level, from B1 to A2	$\sigma^2_{u_4}$	0.00	-	-
	Trend, phase A2	$\sigma^2_{u_5}$	0.00	-	-
	Treatment effect, second AB pair	$\sigma^2_{u_6}$	0.00	-	-
	Treatment effect on trend, second AB pair	$\sigma^2_{u_7}$	0.00	-	-
	Residual variance, baseline	$\sigma^2_{e,A}$	6.08*	0.88	< .001
	Residual variance, treatment	$\sigma^2_{e,B}$	2.69*	0.35	< .001
	Autocorrelation, baseline	$\rho_{(A)}$	0.35*	0.09	< .001
	Autocorrelation, treatment	$\rho_{(B)}$	0.12	0.09	.22

Note. θ_{00} indicates the expected outcome during the baseline phase for class A. $\theta_{00} + \theta_{01}$ indicates the expected outcome during the baseline phase for class B.
* $p < .05$.

7.2.3 Summary of two-level analysis of single-case experimental data

We suggested four plausible models, starting with the most basic two-level model and gradually making it more complex, in order to analyze the dataset of Lambert et al. (2006). By analyzing the data using four different models, we can investigate the extent to which the results are influenced by using different, increasingly complex modeling options. If different results are obtained across models, we recommend single-case researchers to report the different models and discuss the diverse results. The results of the immediate treatment effect estimates using the different models are summarized in Table 7.8 because the single-case researcher is mainly interested in these effects. We can conclude that these results are relatively robust against the different model choices, at least for this empirical illustration. Therefore, our confidence in the conclusion concerning the effectiveness of the treatment is increased. However, if single-case researchers are interested in the variance components estimates, more caution is needed in interpretation because the variance estimates are more sensitive to model choice.

Table 7.8

Summary of Treatment Effect Estimates for Model 1 through Model 4 Using the First Way or the Second Way of Coding the ABAB Reversal Design

		Parameter estimates (*SE*)			
		Model 1	Model 2	[a]Model 3	[a]Model 4
First way of coding	Average shift	-5.40* (0.39)	-5.34* (0.35)		
	Fit statistics				
	-2*log likelihood	1153.4	1124.7		
	AIC	1165.4	1142.7		
	BIC	1166.5	1144.5		
Second way of coding	A1 to B1	-5.66* (0.38)	-5.76* (0.50)	-5.10* (0.69)	-5.53* (0.84)
	A2 to B2	-5.08* (0.49)	-4.92* (0.49)	-5.77* (0.84)	-5.60* (0.84)
	Fit statistics				
	-2*log likelihood	1142.7	1117.6	1106.5	1104.5
	AIC	1170.7	1151.6	1132.5	1132.5
	BIC	1173.4	1150.0	1135.0	1135.3

Note. Standard errors are in parentheses.
[a]Model 3A and Model 4A were not estimated, because the average trend across the two baseline phases and the average trend across the two treatment phases were not of interest.
*$p < .05$.

In the presentation of the four models, we choose to systematically extend the basic two-level model to more complex models. A drawback of this approach is that some of the complexities that have been added may not be needed. An alternative approach, which is illustrated by Singer and Willlet (2003), is to use fit statistics, such as -2 times the log likelihood ratio (i.e., -2LL; Raudenbush & Bryk, 2002), Akaike's information criterion (i.e., AIC; Akaike, 1973), and Bayesian Information Criterion (i.e., BIC; Schwarz, 1978), to choose which complexities to keep (or drop) as the model is being built. The fit statistics for the four models we examined are presented in Table 7.8 and indicate that Model 2 fits the single-case data better (i.e., has smaller values for the fit statistics) than Model 1, but making Model 2 more complex (i.e., Model 3 and Model 4) does not result in better fit statistics. A drawback of using fit statistics to choose a single model is that, with small sample sizes, fit statistics can lead to selection of the incorrect model. Our preference, when working with single-case data, is to estimate treatment effects across a range of plausible models.

Power for testing the treatment effect is also an important issue if small datasets are encountered, which is the case in the two-level analysis of single-cases. In this study, the effects were found to be statistically significant across each of the four models, so power was

adequate for the tests of the average treatment effect. Recently, Ferron et al. (accepted) conducted a simulation study and found that a reasonable power (.80 or higher) was reached when only four participants were included in the study (and the treatment effect equaled a shift of 2 baseline standard deviations). To compare, for 12 participants, a power that exceeded .80 was obtained with effect sizes of one and higher.

In Model 1 through Model 4, continuous outcomes were assumed because the continuous outcome multilevel model has been more extensively studied in previous research (Ferron et al., 2009; Moeyaert et al., 2013a, 2013b, 2013c; Owens & Ferron, 2012; Ugille et al., 2012, 2013; Van den Noortgate & Onghena, 2003a, 2003b, 2008). However, in Lambert et al. (2006), the outcome variable is a count (i.e. per session, the participating students were each observed during 10 intervals, and the number of intervals in which disruptive behavior was observed was recorded). Therefore, we will briefly discuss a basic logistic regression model to show that the multilevel model can be adjusted to model count data. However, further research is needed to investigate how the basic logistic regression model functions for single-case experimental data. Equations 7.6 and 7.7 display the logistic models and $\hat{\varphi}_{ij}$ indicates the expected proportion of trials within session *i* for subject *j* in which the behavior was exhibited: $\hat{\varphi}_{ij} = y_{ij}/10$.

Level 1 (Logistic Model A):

$$log\left(\frac{\hat{\varphi}_{ij}}{1-\hat{\varphi}_{ij}}\right) = \beta_{0j} + \beta_{1j}Phase_{ij} \tag{7.6}$$

Level 1 (Logistic Model B):

$$log\left(\frac{\hat{\varphi}_{ij}}{1-\hat{\varphi}_{ij}}\right) = \beta_{0j} + \beta_{1j}A1B1_{ij} + \beta_{2j}B1A2_{ij} + \beta_{3j}A2B2_{ij} \tag{7.7}$$

When using Equation 7.6 and Equation 7.7, the parameter estimates are expressed on a logit scale, which complicates interpretation. Therefore, we back-transformed the parameter estimates as displayed in Table 7.9: y_{ij} can be calculated by solving the following equation: $\log[\hat{\varphi}_{ij}(1-\hat{\varphi}_{ij}] = \log[\frac{y_{ij}}{10}/(1-\frac{y_{ij}}{10})] = 0.79$. As a consequence, $\frac{y_{ij}}{10}/(1-\frac{y_{ij}}{10}) = \exp(0.82)$, and y_{ij} equals 6.88, indicating the predicted number of challenging behaviors during the baseline phase. The expected logit during the treatment phase equals -1.93. If we back-transform this value, we obtain an average number of challenging behaviors of 1.27 during the treatment phase. By back-transforming the predicted baseline level (i.e., 6.88) and the predicted outcome score during the treatment phase (i.e., 1.27), a treatment effect of -5.61 is

found (= 6.88 – 1.27), which is what we expected from visual analysis. Also, the treatment effect during both AB pairs is statistically significant, $\hat{\theta}_{10}$ = -5.84, $t(8)$ = -12.66, p = .001, and $\hat{\theta}_{30}$ = –5.39, $t(8)$ = -10.47, p < .001.

Table 7.9

Parameter and Standard Error Estimates Resulting from Estimation of the Logistic Model Using the Lambert et al. (2006) Dataset

	Parameter	Parameter Estimate	*SE*	*p*	Back Transformed
Logistic Model A	Fixed coefficient				
Average baseline level	θ_{00}	0.79*	0.19	.003	6.88*
Average outcome score during the treatment	θ_{10}	-2.72*	0.22	< .001	-5.61*
	Variance component				
Baseline level	$\sigma^2_{u_0}$	0.29	0.16	-	-
Treatment effect	$\sigma^2_{u_1}$	0.32	0.24	-	-
Logistic Model B	Fixed coefficient				
Average baseline level, first AB pair	θ_{00}	0.82*	0.16	.001	6.94*
Average treatment effect, first AB pair	θ_{10}	-2.91*	0.23	< .001	-5.84*
Average change in level, from B1 to A2	θ_{20}	2.82*	0.29	< .001	5.65*
Average treatment effect, second AB pair	θ_{30}	-2.58*	0.25	< .001	-5.39*
	Variance component				
Baseline level, first AB pair	$\sigma^2_{u_0}$	0.17	0.12	-	-
Treatment effect, first AB pair	$\sigma^2_{u_1}$	0.22	0.22	-	-
Change in level, from B1 to A2	$\sigma^2_{u_2}$	0.51	0.36	-	-
Treatment effect, second AB pair	$\sigma^2_{u_3}$	0.36	0.28	-	-

Note. * p < .05.

Note that the results for the fixed effect estimates from the analysis recognizing count outcomes are similar to those obtained by treating the outcomes as continuous. For some datasets and models the difference in results between continuous and count outcome models may be substantial, but for this dataset and model, the differences are small, and as a consequence, our confidence in the conclusion that the treatment was effective is strengthened. Although we presented a variety of plausible two-level models, other models are also possible—for instance, models including a quadratic or other nonlinear trend and models with more predictors We only presented a subset of modeling options that seemed most appropriate for this particular dataset, based on a visual analysis of the data.

7.3 From a Two-Level to a Three-Level Framework

7.3.1 Three-level model

The number of published single-case studies is growing rapidly during the last decade, and therefore there is an increasing interest in meta-analyzing these types of studies in order to estimate average treatment effects. The three-level model can be used to synthesize data across cases and across studies. If we pool several studies together, we can examine the generalizability of the results. The synthesis of the studies can inform policy and important decisions can be made based on these results. Van den Noortgate and Onghena (2008) suggested extending the two-level model to a three-level model by adding an index *k* in Equation 7.1 referring to the study. The outcome score, y_{ijk}, indicates the outcome score at measurement occasion *i* for the j^{th} case from study *k*. For a single-case design with one A-phase and one B-phase, the regression equation at the first level looks as follow:

$$\text{Level 1:} \qquad Y_{ijk} = \beta_{0jk} + \beta_{1jk} Phase_{ijk} + e_{ijk} \quad \text{with} \quad e_{ijk} \sim N(0, \sigma_e^2) \tag{7.8}$$

Note that Equation 7.8 is exactly the same as Equation 7.1 with the only difference that Equation 7.8 has an additional index, *k*, indicating the study. Equation 7.8 takes the hierarchical structure of the data into account when combining single-case studies: measurement occasions, *i* (*i* = 1, 2, …*I*), belong to cases, and cases, *j* (*j* = 1, 2, … *J*), belong to studies, *k* (*k* = 1, 2, … *K*). Equation 7.8 can be used to describe continuous data assuming no trends, a homogeneous within-case variance, and residuals that are independent and normally distributed. At the second level of the three-level model, the two coefficients from the first level are modeled as varying across cases within studies:

$$\text{Level 2:} \qquad \begin{aligned} \beta_{0jk} &= \theta_{00k} + u_{0jk} \\ \beta_{1jk} &= \theta_{10k} + u_{1jk} \end{aligned} \qquad \begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0 u_1} \\ \sigma_{u_1 u_0} & \sigma_{u_1}^2 \end{bmatrix} \right) \tag{7.9}$$

These two equations represent the average baseline level and the average treatment effect across cases within study *k*. Also, the deviation of each particular case from the average study-specific baseline phase level (u_{0jk}) and the average study-specific treatment effect (u_{1jk}) can be estimated. When summarizing single-case results over cases within a study, θ_{10k} and $\sigma_{u_1}^2$ are of particular interest, because they represent the average treatment effect and the extent to which this estimated treatment effect varies across cases within the same study. A researcher can go a step further by meta-analyzing the single-case studies (i.e., combining

the single-case results across studies) in order to estimate the average baseline level and the average treatment effect across cases and across studies. Also, the variation in these estimates between studies might be of interest and can be estimated. At the third level of the model, the variation of the level-2 coefficients from equation 7.9 is modeled as follow:

$$\text{Level 3:} \quad \begin{matrix} \theta_{00k} = \gamma_{000} + v_{00k} \\ \theta_{10k} = \gamma_{100} + v_{10k} \end{matrix} \quad \text{and} \quad \begin{bmatrix} v_{00k} \\ v_{10k} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{v_0} & \sigma_{v_0 v_1} \\ \sigma_{v_1 u_0} & \sigma^2_{v_1} \end{bmatrix} \right) \tag{7.10}$$

The interest of the single-case researcher lies typically in γ_{100} indicating the average estimated treatment effect and $\sigma^2_{v_1}$ referring to the deviation of study k from this average estimated effect.

The three-level model is an extension of the two-level model and has the advantage that more general conclusions can be made and that it increases the examination of external validity. For instance, if a low estimate for the between-study variance of the average treatment effect is found, then there is more evidence that the estimated treatment effect is generalizable. If a large amount of between-study variance is found, predictors can be added to explain this variance and therefore more general conclusions regarding the average estimated treatment effect can be made. Moreover, in addition to average estimates across cases and studies, study-specific and case-specific estimates can be obtained using the command "solution" after the random statements. This three-level model takes the hierarchical structure of the data into account: measurements are nested within cases and cases in turn are nested in studies. Also, the between-case and between-study variance of these effects can be estimated. The other advantages of this three-level approach are similar to the two-level approach and include, amongst others, the possibility of modeling different types of trajectories (e.g., nonlinear trends), modeling count data, including autocorrelation, and adding predictors at the three levels (e.g., study quality at the third level and participant-specific predictors such as age at the second level). Also dependent, non-normal and heterogeneous error variances at the three levels can be taken into account.

There has been some work devoted to the empirical validation of the basic three-level models using Monte Carlo simulation studies. The three-level model used to analyze multiple-baseline designs with only a treatment effect has been studied for unstandardized raw data (Owens & Ferron, 2010), and the three-level model applied to unstandardized and standardized multiple-baseline data modeling trends during baseline and treatment phases has also been studied (Moeyaert et al., 2013a, 2013b). These studies indicate that the three-level

model results in unbiased average treatment effect estimates (even if there are a small group of cases and studies included) and that the between-case and between-study variance estimates can be biased if there are a small number of studies ($\leq$ 10) and a small number of cases ($\leq$3 per study) included. Furthermore, this three-level model including time trends has been adapted to model external event effects (Moeyaert et al., 2013c). The misspecification of the error variances at the first level of the three-level model has been investigated by Petit-Bois, Baek, and Ferron (2013) as well as the misspecification of the variance matrix at the second and third level (Moeyaert et al., 2014a).

7.3.2 Empirical illustration of the three-level model

In this section, we illustrate this three-level approach in order to summarize results in seven studies measuring the same outcome variable, namely the effects of pivotal response training with children with autism (measured as the percentage of trials with appropriate speech). In these seven studies, a multiple-baseline across participants design is used. We are mainly interested in the fixed effect estimates (i.e., average baseline level and treatment effect across cases and studies), but we also illustrate that the between-case and the between-study variance can be estimated. Again, the multilevel model is very flexible, and several models can be investigated when analyzing single-case data. We will present four plausible three-level models based on a visual analysis of the data, but other models are also possible. We propose a variety of models to illustrate several modeling options and emphasize that there is no single superior model that works with all three-level single-case datasets. The analysis of multiple-baseline design data is simpler than the analysis of reversal design data in that there is only one transition from the baseline to the treatment phase.

All of the seven multiple-baseline studies we want to combine are characterized by multiple dependent variables (Koegel, Symon, & Koegel, 2002; Koegel, Camarata, Valdez-Menchaca, & Koegel, 1998; Laski et al., 1988; Leblanc, Geiger, Sautter, & Sidener, 2007; Schreibman, Stahmber, Bartlett, & Dufek, 2009; Sherer & Schreibman, 2005; Thorp, Stahmer, & Schreibman, 1995). From these seven studies, we choose to only include the dependent variable measuring appropriate or spontaneous speech. Furthermore, the outcome scale of two of the seven studies was a count (Koegel, Camarata, & Koegel, 1998; Koegel, Symon, & Koegel, 2002) and differed from the outcome scale from the other studies, which was a percentage (the percent of intervals in which the desired behavior occurred). Therefore, we choose to reduce the dataset by only combining results from the five studies in which the

appropriate or spontaneous behavior was measured on the same scale (as a percentage instead of a count). The SAS code is included in Addendum A6.

We start with discussing the results using the basic three-level model in which there is only a shift in level; there are no trends; the errors at the three levels are independent, identically, and normal distributed; and there are no predictors at the higher levels of the model. This basic three-level model will then be modified in several ways in the second part, representing more complex and probably more realistic modeling assumptions. When discussing the results, we choose .05 as the alpha-level. We used SAS 9.3 to conduct the analyses and the SAS code for the basic three-level model (i.e., Model 1) as well as the extensions to this model (Model 2 to Model 4) are contained in Addendum A5.

7.3.2.1 Model 1: basic three-level model

We start by presenting the results obtained by using the most basic three-level model to combine results from the five multiple-baseline design studies. In this basic model, we make a lot of assumptions: there are no time trends, there are no predictors at the second and third level, and the errors at the three levels are independent, identically, and normal distributed (see Equations 7.8 - 7.10). The average baseline level (i.e., γ_{000}) and treatment effect (i.e., γ_{100}) are estimated across cases and across studies in addition to the between-case (co)variance and between-study (co)variance of these estimates. Results are displayed in Table 7.10.

Table 7.10

Parameter and Standard Error Estimates Resulting from Estimation of the Three-Level Analysis of Model 1

	Parameter	Parameter estimate	*SE*	*p*
	Fixed coefficient			
Average baseline level	γ_{000}	19.36*	5.75	.016
Average treatment effect	γ_{100}	31.07*	8.15	.016
	(Co)variance component			
Between-study (co)variance				
Baseline level	$\sigma^2_{v_0}$	96.76	95.31	.155
Treatment effect	$\sigma^2_{v_1}$	271.96	223.79	.112
Covariance between baseline level and treatment effect	$\sigma_{v_0 v_1}$	144.26	124.38	.246
Between-case (co)variance				
Baseline level	$\sigma^2_{u_0}$	316.15*	103.52	.001
Treatment effect	$\sigma^2_{u_1}$	224.11*	83.56	.004
Covariance between baseline level and treatment effect	$\sigma_{u_0 u_1}$	-49.24	70.99	.488
Residual variance	σ^2_e	328.72*	15.70	< .001

Note. * $p < .05$.

From this basic three-level analysis, we conclude that there is a statistically significant average treatment effect that equals 31.07, $t(4.38) = 3.81$, $p = .016$. The variance in the estimated treatment effect between studies is not statistically significant, whereas the variance in this estimated treatment effect between cases is significant and equals 224.11, $Z = 2.68$, $p = .004$. There is also a significant within-case variance.

We suggest three alternatives to Model 1 (i.e. the basic three-level model), based on the graphical presentation of the data in the primary studies, but other models are also possible dependent on the research interest and the specific meta-analysis you want to conduct. In Model 2, we make a less strong assumption by modeling dependence between the residuals at the first level (i.e., autocorrelation) and assuming that the within-case residuals are not necessarily identically distributed across the two phases. In Model 3, we suggest including a time trend in the treatment phase, because the visual inspection of the five primary studies indicates that there is no trend during the baseline but a slightly positive linear trend during the treatment phase. In the last model (i.e., Model 4), we will explore whether predictors at the higher levels of the multilevel model have a significant effect on the estimated outcome scores and if these predictors succeed in reducing the between-case variance, the between-study variance, or both.

7.3.2.2 Model 2

In this first alternative model, we model autocorrelation because in a single-case design, the cases are measured repeatedly, usually with small time periods in between the consecutive measurement occasions. Therefore, it is likely that measurement occasions closer in time are more related than measurements further away in time. In this second model, we also model heterogeneous within-case phase variances. Looking at the primary multiple-baseline studies, we expect that the scores within the baseline phase are more stable in comparison to the scores in the treatment phase. The results of the three-level analysis taking autocorrelation into account and modeling heterogeneous within-case phase variances are presented in Table 7.11.

Table 7.11

Parameter and Standard Error Estimates Resulting from Estimation of the Three-Level Analysis of Model 2

	Parameter	Parameter estimate	*SE*	*p*
	Fixed coefficient			
Average baseline level	γ_{000}	18.72*	5.69	.017
Average treatment effect	γ_{100}	30.50*	7.92	.016
	(Co)variance component			
Between-study (co)variance				
Baseline level	$\sigma^2_{v_0}$	93.93	92.19	.154
Treatment effect	$\sigma^2_{v_1}$	242.43	211.83	.126
Covariance between baseline level and treatment effect	$\sigma_{v_0 v_1}$	125.16	116.82	.284
Between-case (co)variance				
Baseline level	$\sigma^2_{u_0}$	302.95*	100.83	.001
Treatment effect	$\sigma^2_{u_1}$	160.79*	85.61	.030
Covariance between baseline level and treatment effect	$\sigma_{u_0 u_1}$	-11.32	72.47	.876
Residual variance, baseline	$\sigma^2_{e,A}$	167.68*	16.56	< .001
Residual variance, treatment	$\sigma^2_{e,B}$	534.84*	55.88	< .001
Autocorrelation, baseline	$\rho_{(A)}$	0.46*	0.05	< .001
Autocorrelation, treatment	$\rho_{(B)}$	0.60*	0.04	< .001

Note. * $p < .05$.

Similar to the basic three-level model, we found a significant estimated treatment effect: $\hat{\gamma}_{100}$ = 30.50, $t(4.29) = 3.85$, $p = .016$. For the estimated variances, we conclude that the between-case variance estimates are statistically significant and that the between-study variance estimates are smaller in comparison to the ones obtained by the basic three-level model and are not statistically significant. As expected, we found that the variance estimate within the treatment phase is larger (3.19 times) than the estimated variance within the baseline phase. Another important finding is that we found significant autocorrelation both in the baseline and the treatment phase. In the baseline phase, the autocorrelation equals 0.46, $Z = 8.90$, $p < .001$, and in the treatment phase, the measurements closer in time are more related to each other than in the baseline phase: autocorrelation = 0.60, $Z = 14.47$, $p < .001$.

7.3.2.3 Model 3

In this third model, we add a time predictor in the model in addition to modeling autocorrelation and heterogeneous within-case phase variances. For simplicity, we chose to not estimate covariance between regression coefficients at the second and third level. The visual analysis of data from the primary studies indicate relatively stable outcome scores during the baseline phase but slightly increasing outcome scores over time during the treatment phase. Therefore, we modified the level 1 equation by adding time as predictor in the treatment phase:

Level1 $$Y_{ijk} = \beta_{0jk} + \beta_{1jk}Phase_{ijk} + \beta_{2jk}T'_{ijk}Phase_{ijk} + e_{ijk} \quad \text{and} \quad e_{ijk} \sim N(0, \sigma_e^2) \tag{7.11}$$

We indicate in Equation 7.11 that the trend over time predictor, modeled during the treatment phase, is centered around the first measurement occasion of the treatment phase by *T'* (see Figure 7.4). In this way, β_{0jk} represents the average outcome score for case *j* of study *k* during the baseline, and β_{1jk} and β_{2jk} indicate respectively the estimated immediate treatment effect (i.e., the shift in level at the time of the first treatment phase observation) and the time trend during the treatment phase, which are of particular interest. The level-1 coefficients vary at the second level and the third level and allows us to estimate the average treatment effect across studies and the between-case and between-study variance as presented in Table 7.12.

Table 7.12

Parameter and Standard Error Estimates Resulting from Estimation of the Three-Level Analysis of Model 3

Model 3	Parameter	Parameter Estimate	*SE*	*p*
	Fixed coefficient			
Average baseline level	γ_{000}	17.31*	5.74	.024
Average immediate treatment effect	γ_{100}	25.26	10.28	.058
Average trend during treatment	γ_{200}	0.76*	0.15	.019
	Variance component			
Between-study variance				
Baseline level	$\sigma^2_{v_0}$	98.48	93.24	.145
Immediate treatment effect	$\sigma^2_{v_1}$	478.03	334.26	.073
Treatment effect on trend	$\sigma^2_{v_2}$	0.00	-	-
Between-case variance				
Baseline level	$\sigma^2_{u_0}$	277.37*	89.47	.001
Immediate treatment effect	$\sigma^2_{u_1}$	85.82*	42.97	.023
Treatment effect on trend	$\sigma^2_{u_2}$	0.09	0.13	.234
Residual variance baseline	$\sigma^2_{e,A}$	170.01*	16.93	< .001
Residual variance treatment	$\sigma^2_{e,B}$	268.58*	19.19	< .001
Autocorrelation, baseline	$\rho_{(A)}$	0.47*	0.05	< .001
Autocorrelation, treatment	$\rho_{(B)}$	0.21*	0.05	< .001

Note. * $p < .05$.

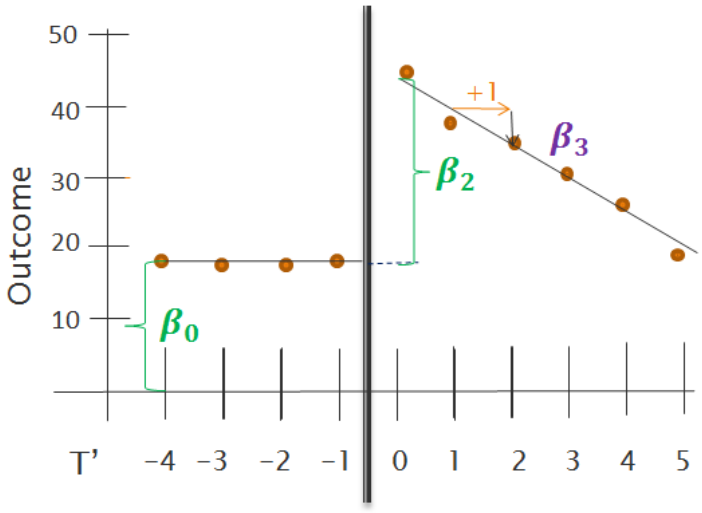

Figure 7.4. Graphical presentation of the coefficients in Equation 7.11 based on hypothetical AB design data.

An interesting finding is that the estimated immediate treatment effect is not statistically significant when a time trend during the treatment phase is modeled and equals 25.26, $t(4.98) = 2.46$, $p = .058$. The estimated time trend, $\hat{\gamma}_{200}$, equals 0.76 and is statistically significant, $t(2.8) = 4.92$, $p = .019$. The estimated between-case variance estimates are significant, except for the trend during the treatment. Also the estimated residual variances are significant. Notwithstanding the modeling of a time trend, the estimated autocorrelation within both the baseline and treatment phase remain positive and statistically significant: autocorrelation during the baseline and the treatment phase equal 0.47, $Z = 9.05$, $p < .001$, and 0.21, $Z = 4.06$, $p < .001$, respectively. However, the estimated autocorrelation during the treatment phase is smaller in comparison to the estimated autocorrelation modeled in Model 2. Note that we chose to center the time predictor around the first measurement occasion of the treatment phase because we wanted to estimate the difference in outcome score between the baseline data and the treatment data at the first measurement occasion in the treatment. A single-case researcher might be interested in the difference in outcome scores at another later point in time, for instance at the third measurement occasion in the treatment. In this case, the time variable has to be centered around that value. If we center time around the middle measurement occasion of the treatment phase, then we would obtain an estimated treatment effect that is more similar to the average shift in level from the previous models.

7.3.2.4 Model 4

In the previous models, we only added predictors at the first level of the three-level model. However, when combining data over cases and over studies case-specific and study-specific characteristics can be included in order to explain between-case and between-study variability in estimated effects. Age, expressed in years, is a case-specific characteristic that was coded in the primary studies and can be included as a second-level predictor. We expect that this predictor will influence the estimated baseline level and that the estimated treatment effect is independent of age. In order to conduct a meaningful analysis, we centered age around the average age, because otherwise the variation in intercept is estimated for participants having an age of zero. We also model autocorrelation and heterogeneous within-case phase variances. Similar to the previous model, we chose to not estimate covariance between regression coefficients at the second and third level for simplicity. The level two and level three equations look as follow:

Level 2
$$\begin{aligned}\beta_{0jk} &= \theta_{00k} + \theta_{01k} age_{jk} + u_{0jk} \\ \beta_{1jk} &= \theta_{10k} + u_{1jk} \\ \beta_{2jk} &= \theta_{20k} + u_{2jk}\end{aligned} \quad \text{and} \begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{2jk} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0u_1} & \sigma_{u_0u_2} \\ \sigma_{u_0u_1} & \sigma^2_{u_1} & \sigma_{u_1u_2} \\ \sigma_{u_0u_2} & \sigma_{u_1u_2} & \sigma^2_{u_2} \end{bmatrix} \right) \tag{7.12}$$

Level 3
$$\begin{aligned}\theta_{00k} &= \gamma_{000} + v_{00k} \\ \theta_{10k} &= \gamma_{100} + v_{10k} \\ \theta_{20k} &= \gamma_{200} + v_{20k}\end{aligned} \quad \text{and} \begin{bmatrix} v_{0jk} \\ v_{1jk} \\ v_{2jk} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{v_0} & \sigma_{v_0v_1} & \sigma_{v_0v_2} \\ \sigma_{v_0v_1} & \sigma^2_{v_1} & \sigma_{v_1v_2} \\ \sigma_{v_0v_2} & \sigma_{v_1v_2} & \sigma^2_{v_2} \end{bmatrix} \right) \tag{7.13}$$

The results of the fixed effect estimates and variance components are displayed in Table 7.13.

Table 7.13

Parameter and Standard Error Estimates Resulting from Estimation of the Three-Level Analysis of Model 4

Model 4	Parameter	Parameter estimate	*SE*	*p*
	Fixed coefficient			
Average baseline level	γ_{000}	17.57*	5.90	.028
Average effect of predictor age during baseline	γ_{010}	-0.05	0.31	.864
Average treatment effect	γ_{100}	25.28	10.30	.058
Average trend during treatment	γ_{200}	0.76*	0.15	.019
	Variance component			
Between-study variance				
Baseline level	$\sigma^2_{v_0}$	97.56	92.40	.146
Treatment effect	$\sigma^2_{v_1}$	488.88	335.61	.073
Trend during treatment	$\sigma^2_{v_2}$	0.00	-	-
Between-case variance				
Baseline level	$\sigma^2_{u_0}$	277.40*	89.45	.001
Treatment effect	$\sigma^2_{u_1}$	85.92*	43.00	.023
Trend during treatment	$\sigma^2_{u_2}$	0.09	0.13	.234
Residual variance baseline	$\sigma^2_{e,A}$	170.00*	16.93	< .001
Residual variance treatment	$\sigma^2_{e,B}$	268.57*	19.19	< .001
Autocorrelation, baseline	$\rho_{(A)}$	0.47*	0.05	< .001
Autocorrelation, treatment	$\rho_{(B)}$	0.21*	0.05	< .001

Note. * $p < .05$.

The average immediate effect of treatment was estimated to be 25.28, $t(4.97) = 2.46$, $p = .058$. The estimated effect of the predictor age on the baseline level is not statistically significant and equals –0.05, $t(8.71) = –0.18$, $p = .864$. The negative value of the predictor age means that the older the case is, the lower the estimated baseline level. The ages of the participants included in this three-level analysis ranged from 2 to 57. The estimated baseline level for a participant with age 2 equals 17.57 + (–0.05 * 2) = 17.47, whereas the estimated baseline level equals 17.57 + (–0.05 * 57) = 14.71 for a participant with age 57. Furthermore, the between-case variances of the intercept and the immediate treatment effect and the within-case variances are statistically significant matching findings from previous models. The autocorrelation during the baseline and the treatment phase is statistically significant and equals .47, $Z = 9.04$, $p < .001$ and .21, $Z = 4.06$, $p < .001$, respectively.

7.3.3 Summary of three-level analysis of single-case experimental data

Table 7.14 provides a summary of the immediate treatment effect estimates for each proposed model (Models 1 through 4).

Table 7.14

Summary of Treatment Effect Estimates for Model 1 through Model 4 Using the First Way or the Second Way of Coding the ABAB Reversal Design

	Parameter estimates (standard errors)			
	Model 1	Model 2	Model 3	Model 4
Average (immediate) treatment effect	31.08* (5.75)	30.50 * (7.92)	25.26 (10.28)	25.28 (10.30)
Fit statistics				
-2*log likelihood	8181.0	7798.8	7678.0	7678.0
AIC	8199.0	7822.8	7702.0	7704.0
BIC	8195.5	7798.8	7678.0	7678.0

Note. Standard errors are in parentheses.
*$p < .05$

Notwithstanding each suggested model has its own assumptions, we found similar results for the treatment effect estimates across Model 1 and Model 2 on the one hand and Model 3 and Model 4 on the other hand. The reason is that, in Model 3 and Model 4, a time trend during the treatment phase is modeled, which is not the case in Models 1 and 2. The positive estimated time trend during the treatment phase in Model 3 and 4 resulted in an estimated outcome score at the start of the treatment phase that was lower in comparison to Models 1 and 2. If single-case researchers are interested in the variance components estimates, more caution is needed when choosing the analysis model, because the variance estimates depend more on the selected model. Thus, interpretation of variance estimates from the models estimated here should be made with some caution. If different results are obtained across models, we recommend that single-case researchers report the different models and discuss the diverse results. Note that for all suggested models, the estimated standard errors at the study level are larger than the estimated standard errors at the second level which has consequences for the significance testing. For instance, the between-study variance of the immediate treatment is large but found to be not statistically significant because of the large estimated standard error (i.e., the estimated standard error is large in comparison to the parameter estimate, resulting in a small *t*-statistic and a large *p*-value), whereas the smaller estimated between-case variance of the immediate treatment effect estimate is found to be statistically significant. There are no problems concerning the power of detecting the

treatment effect, as a total of 27 participants (spread over 5 studies) are included in this study (Moeyaert et al., 2013a). Also, fit statistics (i.e., -2LL, AIC, and BIC) are presented in Table 7.14 and indicate that Model 2 fits the SSED data better than Model 1 (i.e., has smaller values for the fit statistics). Model 3 and 4 fit the data better than Model 1 and 2, but the difference between Model 3 and 4 is negligible.

Although we demonstrated a variety of different model extensions, other extensions are also possible—for instance adding quadratic or other nonlinear trends, adding more predictors, or adding covariance between the random effects at the different levels. We only presented the modeling options that are most plausible, based on visual analysis of the primary studies. The four models are presented in an increasing level of complexity. However, single-case researchers may choose another way of modeling building, such as the approach suggested by Singer and Willet (2003) in which nonsignificant parameters are removed.

7.4 Discussion

Using the multilevel model (either the two-level or three-level model) to summarize single-case results over cases, over studies, or both has multiple advantages. Multilevel models can provide detailed information regarding the treatment effects (e.g., estimates of case-specific immediate treatment effects, case-specific trend shifts, level shifts across cases and across studies, average trend shifts across cases and across studies, and variance in effects across participants and studies). The multilevel models can be adapted for different designs (e.g., multiple-baseline, reversal, and alternating treatments designs) and for different types of outcomes (e.g., continuous, binary, and count), while also taking into account trends, autocorrelation, heterogeneity, and nesting of cases within studies. To show the flexibility of the multilevel model, we suggested a variety of plausible two-level and three-level models in this article, and we provided empirical illustrations and interpretation of results.

In the first part of the study, we presented the two-level analysis of single-case studies using two different ways of coding data based on the ABAB design. We combined the data of nine replicated ABAB designs using the basic two-level model and proposed several alternatives. The results of the fixed effects estimates were relatively robust against several modeling options. However, variance estimates varied, but we have to interpret their values with caution because previous simulation studies have indicated that variance estimates can be biased, especially when a small number of measurement occasions and cases are involved.

In this first part, there were enough measurement occasions, but the number of cases is likely too small for valid inferences about variance values. Ferron et al. (2009) do not encourage interpreting the variances if there are eight or fewer cases due to the bias that they found. The current study included nine cases, but combining more than eight cases has not yet been examined. Moreover, the study of Ferron et al. (2009) focused on multiple-baseline designs, whereas in this study, nine replicated ABAB designs were combined.

In the second part of the study, we focused on three-level analyses, combining single-case results over cases and over studies. The three-level analyses of Owens and Ferron (2012), estimating the treatment over cases and over studies, showed that the estimate of the average baseline level and average treatment effect lead to unbiased estimates; however, the estimates of the variance components (between-case and between-study variance) are questionable. Similar conclusion were obtained by Moeyaert et al. (2013a, 2013b) in which trends were included. So the results of the variance estimates in the second part of the current study have to be interpreted with some caution.

We only presented a limited number of plausible two-level and three-level models, but others are also possible. For instance, in this study, we choose to model amongst others autocorrelation, heterogeneous within-case variance, and trends during the treatment phase. However, other modeling options such as trends during the baseline phase and different types of predictors are also possible. Estimating multiple models has two purposes. First it allows us to illustrate the flexibility of the multilevel model and to illustrate how convenient it is to adjust the model according to the assumptions one makes and according to the researcher's interests. Second, it illustrates the practice of estimating multiple alternative models. Any model rests on a series of assumptions and the amount of data available to single-case researchers is often not sufficient to rigorously test and validate these assumptions. As a consequence, the assumptions and model can be questioned, leading to uncertainty in the conclusions reached. By considering a range of plausible models and assumptions, researchers can determine the degree to which the effect estimates and conclusions are sensitive to the specific assumptions made. If the same conclusions are reached across a range of plausible assumptions, confidence in the conclusions can be enhanced. We advise researchers not to focus on one model but to conduct multiple plausible multilevel analyses and investigate whether the results depend on the modeling choices.

In this study, significant treatment effect estimates across cases (two-level analysis) and across cases and studies (three-level analysis) are found. However, this does not imply significant treatment effect for all cases included in the two-level or three-level analysis. The

multilevel analysis does not throw away information about these individual cases. On the contrary, it allows estimating and explaining differences between individual cases and obtaining case-specific treatment effect estimates by using empirical Bayes techniques.

The parameters were estimated and tested using the maximum likelihood (ML) estimation in SAS. However, ML estimation of multilevel models is also included in HLM, MLwiN, R, SPSS, and Stata. Previous simulation studies indicate that using ML (similar to using restricted maximum likelihood), which is based on large sample theory, to estimate multilevel models for single-case data leads to biased variance estimates, especially with a smaller number of units at level 2 or level 3 (Ferron et al., 2009; Moeyaert et al., 2013a, 2013b; Owens & Ferron, 2012). Alternatives, which may result in less biased variance estimates in small samples, are Bayesian estimation (Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008) and bootstrapping (Wang, Xie, & Fisher, 2012) procedures. Further research is needed assessing use of these alternative procedures.

We illustrated the multilevel approach using the raw data, but it is also possible to synthesize the data at the first level using effect sizes. These effect sizes can then be combined over cases and over studies using a multilevel meta-analysis instead of multilevel analysis. Originally, Van den Noortgate and Onghena (2008) proposed regression coefficients as effect size estimator, and Ugille et al. (2012) conducted a simulation study to empirically validate the multilevel meta-analysis of this effect size estimator. Similar to the effect size estimator presented by Hedges, Pustejovsky, and Shadish (2012), namely the standardized mean difference, the effect size proposed by Van den Noortgate and Onghena (2008) can be converted to an effect size that can be used in meta-analysis of both single-case experimental data and group-comparison designs. For more details about this effect size, we refer to the article of Van den Noortgate and Onghena (2008).

The studies combined in the three-level analysis were chosen on purpose to make sure that the outcome variable was measured on the same scale. If studies are not on the same scale, we advise researchers to first standardize the single-data before combining them in a multilevel analysis. The standardization method for continuous outcomes was introduced by Van den Noortgate and Onghena (2008). They proposed performing an ordinary least squares regression for each subject from one study separately (for instance, using Equation 7.8) in order to estimate the residual within-subject standard deviation ($\hat{\sigma}_{ejk}$). Thereafter, the individual scores ($Y_{ijk}{}'s$) are divided by the estimated residual within-subject standard deviation ($\hat{\sigma}_{ejk}$). The residual within-subject standard deviation estimate reflects the

differences in how the dependent variable is measured, and thus dividing the original raw scores in a study by this variability provides a method of standardizing the scores. The standardized scores can then be used in the multilevel model. This standardization method in contexts of the three-level modeling of continuous single-case data has been explored and studied. Moeyaert et al. (2013b) found that the standardization procedure resulted in more biased and less precise treatment effect estimates and that these problems became negligible as series lengths increased (i.e., larger than 20).

In this article, we discussed several plausible multilevel models for the analysis of two-level and three-level single-case data. Although we selected a variety of different modeling options based on visual inspection of the data, other options are also possible. We will never know what the correct underlying model is, but by showing that the results of interest are similar across a range of plausible models, confidence in the obtained findings can be increased. Applied researchers are thus encouraged to explore several multilevel models to analyze their data based on visual analysis of the primary studies and their research interests.

Chapter 8|

Estimating Intervention Effects Across Different Types of Single-Subject Experimental Designs: Empirical Illustration[7]

Abstract

The purpose of this study is to illustrate the multilevel meta-analysis of results from single-subject experimental designs of different types, including AB phase designs, multiple-baseline designs, ABAB reversal designs, and alternating treatment designs. Current methodological work on the meta-analysis of single-subject experimental designs often focuses on combining simple AB phase designs or multiple-baseline designs. We discuss the estimation of the average intervention effect estimate across different types of single-subject experimental designs using several multilevel meta-analytic models. We illustrate the different models using a re-analysis of a meta-analysis of single-subject experimental designs (Heyvaert, Saenen, Maes, & Onghena, 2014). The intervention effect estimates using univariate three-level models differ from those obtained using a multivariate three-level model that takes the dependence between effect sizes into account. Because different results are obtained and the multivariate model has multiple advantages, including more information and smaller standard errors, we recommend researchers to use the multivariate multilevel model to meta-analyze studies that utilize different single-subject designs.

Keywords: single-subject experimental design, univariate multilevel modeling, multivariate multilevel modeling, average intervention effect

[7] This chapter has been published as*:* Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M., Beretvas, S.N.,& Van den Noortgate, W. (2014c). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly*, 52 (2).

8.1 Introduction

To help improve schools and the care for students within schools it is critical to understand the effects of interventions, the degree to which those effects vary across students and settings, and the characteristics of the students and settings that are associated with intervention effectiveness. As a consequence, it is important to not only conduct intervention studies that focus on individuals, but also to maximize what is collectively learned through the synthesis of this intervention research. In school psychology, single-subject experimental designs (SSEDs) are often used (Kratochwill, 1985; Wacker, Steege, & Berg, 1988) to evaluate the effect of the experimental manipulation of an independent variable (e.g., a specific intervention) on the dependent variable (e.g., outcome scores). The focus of the study is then on a single entity, for instance, on a single student or a single teacher-student dyad, or a small number of entities (e.g., McGoey & DuPaul, 2000; Briesch & Chafouleas, 2009). In an SSED, the independent variable is manipulated by the experimenter, and the dependent variable is measured repeatedly for this entity under different levels of the independent variable. Reasons for the popularity of SSEDs in school psychology research include the focus on the individual that parallels the care for the individual in applied settings and the fact that an SSED is one of the only eligible design options if rare or unique conditions are involved. Furthermore, an SSED is warranted when researchers are interested in within-subject variability, rather than the between-subject variability. Although SSED studies are commonly undertaken, techniques used for the synthesis of SSED studies has not received the same level of attention as the synthesis of group studies. In this manuscript, we focus on techniques for the meta-analysis of results from multiple SSED studies, allowing for the exploration of the generalizability of the intervention effects and the study of moderator variables that might influence an intervention's effects. Assessment and improvement of techniques to synthesize SSED results is of the utmost importance, as the number of published SSEDs is increasing at an astonishing rate (Social Science Citation Index).

Visual analysis has been the traditional method for evaluating treatment effects in SSED research, but is by itself less suited for synthesizing literature in an objective way (Manolov & Solanas, 2013). The evidence-based movement in SSED context has emphasized the need for quantitative summaries of studies' results, especially for making them available for meta-analytic purposes (Jenson et al., 2007). Parker and Brossart (2003) state three major advantages of developing effect sizes. First, effect sizes express the magnitude of the relationship between an intervention and outcome. Second, effect sizes provide a continuous

(rather than categorical) evaluation of treatment success. Third, effect sizes are not systematically affected by sample size. Manolov and Solanas (2013) argue that statistical and visual analysis are complementary because of the lack of formal decision rules in visual analysis, the corresponding lack of objective and replicable outcomes, the idea that practitioners would have more confidence in treatment effectiveness when visual and statistical analysis inferences coincide, and the increased credibility of SSED findings for the scientific community given the use of statistical analysis.

Effect sizes based on non-overlap statistics and those based on regression models have been considered for synthesizing single-case studies (Maggin et al., 2011; Methe et al., 2012). We focus on regression-based effect sizes because they have the capacity to model complex data patterns (Center et al., 1985-1986; Van den Noortgate & Onghena, 2003a). For instance, they can account for non-linear trends and dependent error structures. The regression approach also allows an estimate of the baseline level, the trend during the baseline, an immediate treatment effect and a treatment effect on slope. Also within-phase and within-case variability can be estimated. For a more in depth discussion of regression-based effect sizes and their combination across SSED cases and studies using multilevel regression models, we refer the reader to Van den Noortgate and Onghena (2003a, 2003b).

Regression-based effect size estimates can be combined across cases and across studies using the multilevel model, which is an extension of the single-level regression model (Van den Noortgate & Onghena, 2003a). In recent years, much attention has been paid to the synthesis of simple AB phase designs and multiple-baseline designs using the two-level model (Ferron et al., 2009; Ferron et al., 2010; Van den Noortgate & Onghena, 2003a) or the three-level model (Moeyaert et al, 2013a; Owens & Ferron, 2010; Van den Noortgate & Onghena, 2008). In the two-level model, data are summarized across cases within a single study which implies a two-level structure: measurement occasions are nested within cases in one SSED study. The three-level analysis goes one step further and can be used to combine data across cases and across studies, which includes a three-level structure: the measurement occasions are clustered within cases and cases are clustered within studies. Using the multilevel modeling framework, the following research questions can be resolved: (1) What is the magnitude of the average treatment effect across cases and across studies?; (2) What is the magnitude and direction of the case-specific intervention effect?; (3) How much does the treatment effect vary within cases, across cases and/or across studies? and (4) Does a (case and/or study level) predictor influence the treatment's effect?

While multilevel modeling techniques are proposed to combine data from simple AB phase designs and multiple-baseline designs, combining data from ABAB phase designs and alternating treatment designs has hardly received any attention. However, Shadish and Sullivan (2011) found that 70% of all published SSEDs in 2008 are characterized by a multiple-baseline design, an alternating treatment design or an ABAB phase design. Combining several types of SSEDs is therefore of importance to increase both the validity and credibility of the intervention effect estimates. If the same conclusion is based on a synthesis of results from different types of SSED designs, then there is more confidence that the results are due to the intervention and not to some outside experimental factors (i.e. internal validity). Combining data from different designs can enhance the external validity of the synthesis' findings because they are based on more diverse data. If several SSED studies' results are combined, then data from multiple studies including one or multiple cases are used thereby providing more information and resulting in more precise treatment effect estimates (i.e., smaller standard errors and narrower confidence intervals). Given the potential benefits of combining results from several types of SSEDs, it is surprising that past research focusing on multilevel analysis of SSEDs has not focused on this issue. The synthesis of single-case results across cases and across studies using the multilevel modeling framework is still in its infancy. Previous research focused on the validation of the basic three-level model in order to obtain a better understanding and assessment of the multilevel modeling framework in this context (Moeyaert et al., 2013a; Ugille et al., 2012). In this previous research, suggestions are made to gradually extend this multilevel model in order to represent more realistic situations such as taking different design types into account. Combining several types of SSEDs in a multilevel modeling framework is challenging, because choices have to be made concerning the coding of data from the different SSEDs and also methodological questions arise concerning the dependency between effect sizes. Therefore in this study, we extend the multilevel meta-analytic method proposed by Van den Noortgate and Onghena (2008) to combine several types of SSEDs We start with a review of the basic characteristics of the AB and ABAB phase designs, the multiple-baseline designs and the alternating treatment designs. We continue by presenting the multilevel modeling framework for combining these designs. Models are illustrated using empirical data.

In order to make this study results comparable to the study of Heyvaert et al. (2014), we will assume flat baseline and treatment levels, and homogeneous within-case variance. The main interest lies in the average intervention effect estimate (and not in changes in slope or variability in outcome scores due to the intervention). Note that according to the What Works

Clearinghouse (*WWC*) standards (Kratochwill et al., 2010), a flat baseline is needed prior to intervening and therefore this assumption is not unreasonable. Trends and heterogeneous within-case variance can be modeled using the regression-based effect size estimator as explained and illustrated in more detail in Moeyaert et al. (2014), but these extensions of the basic model are beyond the scope of this manuscript.

8.2 AB Phase Design

An AB phase design is the simplest type of SSED and is characterized by one baseline phase and one treatment phase. The baseline phase is critical as it serves as the basis for predicting the level of performance in the near future if the intervention is not in effect. According to the *WWC* standards (Kratochwill et al., 2010), it is important to observe baseline performance minimally across three measurement occasions to provide a sufficient basis for making a prediction of future performance. Figure 8.1 illustrates how observations during the baseline phase are used to predict future performance.

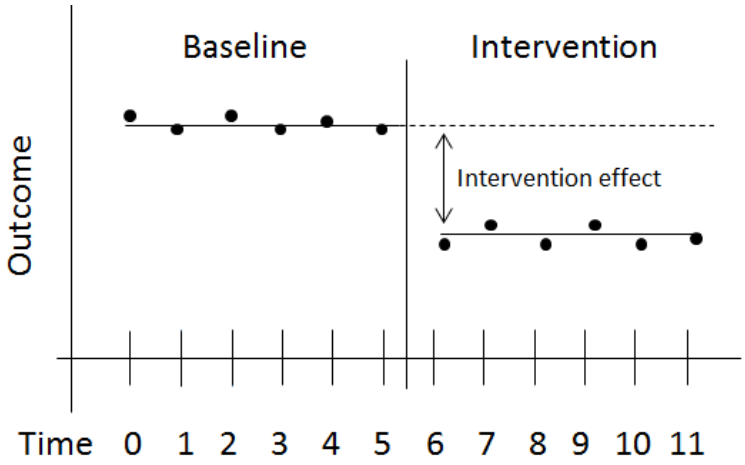

Figure 8.1. Graphical presentation of the intervention effect for an AB phase design. The solid lines refer to the actual outcome level and the dashed lines refer to the projected outcome level.

In order to analyze SSED data as presented in Figure 1, we can apply the following basic regression equation in which the outcome score is regressed on an intercept [i.e., expected outcome level when the independent variable(s) equal(s) zero] and a dummy variable indicating the phase (i.e., the dummy variable, *Intervention*, equals zero if the measurement occasion belongs to the baseline phase, one otherwise).

$$y_i = \beta_0 + \beta_1 Intervention_i + e_i \text{ and } e_i \sim N(0,\ \sigma_e^2) \quad (8.1)$$

Using Equation 8.1, β_1 indicates how performance is increased when going from the baseline to the intervention phase, and therefore can be interpreted as the intervention effect. Research synthesists of SSED data often assume homogeneous, normally distributed and independent errors, such as in Equation 8.1, to facilitate estimation and interpretation of the parameters.

8.3 ABAB Reversal Designs

A possible extension of the AB phase design, in which we distinguish between four phases (i.e., two baseline phases, A1 and A2, and two intervention phases, B1 and B2), is the ABAB reversal design. During each phase, the level of behavior is assessed and is projected to the next phase to predict the level of behavior in the near future (see Figure 8.2). Other variations of reversal designs are possible, but serve the same purpose (e.g., ABABAB, ABCBC, see Barlow et al., 2009 and Kazdin, 2011).

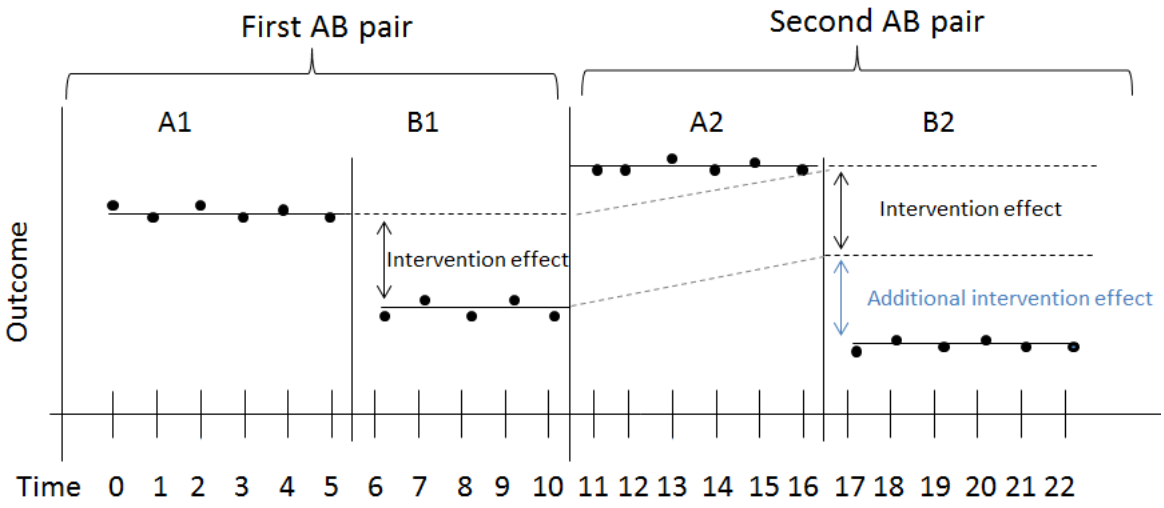

Figure 8.2. Graphical presentation of the intervention effect in the first and the second AB pairs for an ABAB reversal design using hypothetical data. The solid lines refer to the actual outcome level and the dashed lines refer to the projected outcome level.

We discuss the scenario in which stable outcome scores during baseline and intervention phases are obtained. Depending on the research interest, different parameters can be estimated. For instance, one could estimate the intervention effect during the second AB pair as a kind of control for the effect observed during the first AB pair. Single-subject researchers are especially interested in the intervention effect in both first and second AB pairs and more specifically whether the intervention effect in the second AB pair is equal to the one obtained in the first AB pair. Therefore, two dummy variables, namely $Intervention_i$ and $Pair_i$ can be modeled as discussed in Moeyaert et al. (2014c). The dummy, $Intervention_i$, indicates

whether measurement *i* is part of a baseline phase (i.e., A1 or A2) or an intervention phase (i.e., B1 or B2). If the measurement occasion belongs to B1 or B2, then $Intervention_i$ equals one, otherwise zero. The dummy variable, $Pair_i$, indicates whether the measurements belong to the first (Pair = 0) or the second AB pair (Pair = 1). The equation that can be used to analyze an ABAB reversal design, assuming no trends and homogeneous within-subject variance, looks as follows:

$$Y_i = \beta_0 + \beta_1 Intervention_i + \beta_2 Pair_i + \beta_3 Intervention_i Pair_i + e_i \quad (8.2)$$
$$\text{and } e_i \sim N(0, \sigma_e^2)$$

Using Equation 8.2, β_0 and β_1 refer to the baseline level and the intervention effect, respectively, during the first AB pair; $\beta_0 + \beta_2$ indicates the second baseline level and $\beta_1 + \beta_3$ refers to the change in level when the treatment starts in the second AB pair. Using Equation 2 therefore results in two coefficients of particular interest, namely, β_1 and $\beta_1 + \beta_3$. β_1 indicates the intervention effect for the first AB pair and β_3 indicates the difference in the intervention effect of the first AB pair and the second AB pair, which we label the additional intervention effect in Figure 8.2. Other coding options for the dummy variables in ABAB phase designs are also possible and provide alternative interpretation of coefficients. For a detailed discussion, we refer the reader to Moeyaert et al. (2014b).

8.4 Multiple-Baseline Designs

A multiple-baseline design (MBD) is characterized by the simultaneous implementation of an AB phase design to different subjects, behaviors or settings (Ferron & Scott, 2005; Onghena, 2005). In this study, we focus on the multiple-baseline across subjects design. An important feature is that the intervention is introduced to the subjects at different points in time. As in the AB and ABAB phase designs, treatment effects can be evaluated in the MBDs by comparing the level of performance during the intervention and the projected baseline level. Because the MBDs are comprised of simultaneously repeated AB designs across different subjects (as opposed to ABAB designs wherein the AB phases are replicated within a single participant), the MBD will be analyzed as separate AB phase designs using Equation 8.1 for each participant. The MBD design and the intervention effects are presented in Figure 8.3.

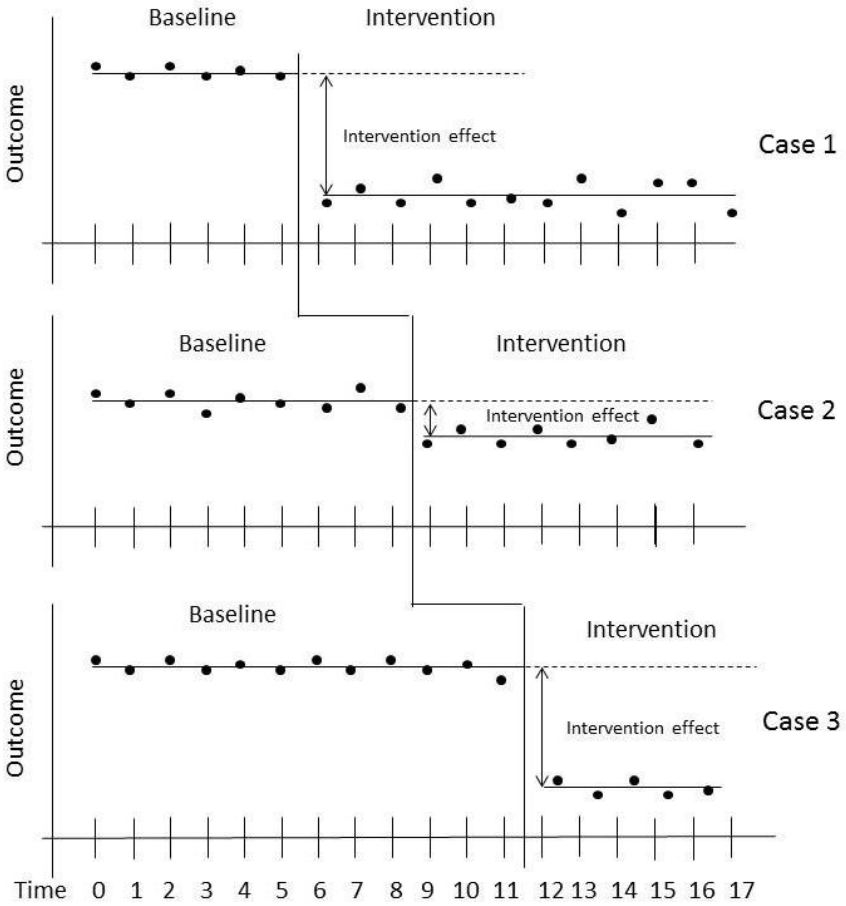

Figure 8.3. Graphical presentation of the intervention effect for an MBD across three subjects using hypothetical data. The solid lines refer to the actual outcome level and the dashed lines refer to the projected outcome levels.

8.5 Alternating Treatment Designs

The difference between alternating treatment designs (ATDs) and previously presented SSEDs is that in ATDs, two or more interventions are rapidly (sometimes randomly) alternated within a single subject. Because we are interested in the estimate of the intervention effect as the difference in outcome scores between intervention and baseline phases, we only discuss the ATDs characterized by multiple interventions during the intervention phase and a baseline phase. The following regression equation can be used in order to analyze data obtained by an ATD:

$$Y_i = \beta_0 + \beta_1 Intervention_{1i} + \beta_2 Intervention_{2i} + e_i \text{ and } \sigma_e^2 \sim N(0, \sigma_e^2) \quad (8.3)$$

The interpretation of the coefficients in Equation 8.3 is straightforward and is depicted in Figure 8.4. $Intervention_{1i}$ and $Intervention_{2i}$ are two dummy coded variables. $Intervention_{1i}$ equals 1 if observation i belongs to the intervention phase, 0 otherwise. $Intervention_{2i}$ equals 1 of observation i belongs to the second treatment, 0 otherwise. As a

consequence, if $Intervention_{1i}$ equals 1 and $Intervention_{2i}$ equals 0, the observation is made during the first intervention, and β_1 can be interpreted as the effect of the first intervention on the outcome score and β_2 as the effect of the second intervention, on top of the effect of the first intervention(i.e., the difference between the first and second intervention effect). For a detailed discussion, we refer the reader to Moeyaert et al. (2014b).

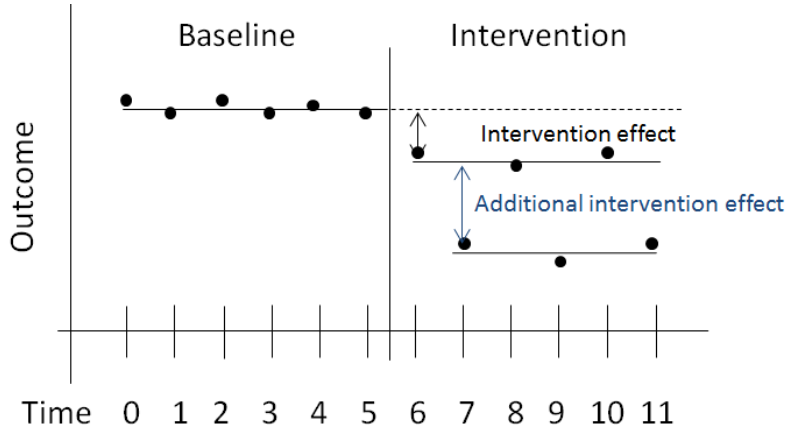

Figure 8.4. Graphical presentation of the intervention effect for an ATD using hypothetical data. The solid lines refer to the actual outcome level and the dashed lines refer to the projected outcome levels.

Equation 8.1 for the AB phase designs and MBDs, Equation 8.2 for the ABAB phase designs and Equation 8.3 for the ATDs all share in common that β_1 refers to the intervention effect either for the first intervention or in the first AB pair. This is fundamental for the remainder of this study. β_2 and β_3 refer to the additional effect of a second intervention (ATDs) and the intervention during the second AB pair (for ABAB phase designs) respectively.

8.6 Three-Level Meta-Analysis Across SSED Types

8.6.1 Effect size

As far back as in the mid-1970s, large numbers of studies, addressing similar underlying research questions, have been published at an astounding rate. Meta-analytic techniques have been proposed to integrate research findings across studies as a basis for examining external validity of treatment effects (Glass, 1976). By pooling results from several studies together, an average treatment effect estimate can be obtained as well as estimates of the variation in the treatment effect across cases and studies. Also, by pooling together results from multiple studies, more reliable treatment effect estimates can be obtained which can inform research and policy. Important advantages of using meta-analysis for SSEDs are that it overcomes the limitations associated with small sample sizes encountered in individual studies, enhances the

power to detect effects of interest, and increases the precision of treatment effect estimates. In addition, differences in intervention effect estimates as a function of individual and study characteristics can be investigated.

In a typical meta-analysis, standardized effect sizes, based on the mean difference in outcome variable between treated and untreated subjects divided by the within-group standard deviation are summarized. In this study, we use another type of effect size, namely regression coefficients as suggested by Van den Noortgate and Onghena (2008). These effect sizes can easily be obtained by conducting an ordinary least square (OLS) regression analysis for each subject using Equations 8.1 to 8.3. Depending on the SSED type and the underlying assumptions, a regression equation is chosen and regression coefficients of interest are estimated (and will serve as the foundation of the effect size in the multilevel analysis). In this study, the intervention effect estimate is represented by b_1 and is an estimate of the true intervention effect, β_1. For the ABAB phase designs and the ATDs, an additional effect size estimate is obtained, namely b_2 or b_3, representing the estimate of the true deviation of the second from the first intervention effect in an ATD (β_2) and the true deviation of the intervention effect in the second AB pair from the intervention effect in the first AB pair for ABAB phase designs (β_3).

8.6.2 Standardized and bias-corrected effect sizes

Once the subject-specific effect sizes, b_1, and possibly (depending on the design) b_2 or b_3 are estimated by conducting an OLS regression analysis for each subject, the next step involves combining the effect estimates across subjects and across studies in order to estimate the average effect size. However, it is reasonable that different studies and different types of studies use a different scale to measure the dependent variable and therefore the effect sizes are not on comparable scales and cannot be combined. For instance, in one study the dependent variable might be measured on a scale from one to ten whereas in another study a scale from one to five might be used. One way to standardize SSED data is by dividing the estimated subject-specific effect size by the estimated residuals' standard deviation as suggested by Van den Noortgate and Onghena (2008). Simulation studies (Moeyaert et al., 2013b; Ugille et al., 2012) show that this works well if the number of observations per subject is larger than 10. The residuals' standard deviation is used as a reflection of the scale used in the original single-subject. b_{ajk} with a = 1, 2 or 3 represents an effect size estimate for subject j from study k. In particular, b_{1jk} indicates the intervention effect estimate for the reference treatment and the first AB pair, b_{2jk} is the estimate of the difference in the effect of the

second versus the first intervention for an ATD, and b_{3jk} is the deviation of the intervention effect estimate during the second versus first AB pair for an ABAB phase design. The standardized effect size, b'_{ajk}, is then obtained by dividing the estimated effect size, b_{ajk}, by the residuals' standard deviation, $\sqrt{\hat{\sigma}_e^2}$ as follows:

$$b'_{ajk} = \frac{b_{ajk}}{\sqrt{\hat{\sigma}_e^2}} \tag{8.4}$$

For more details, we refer the reader to Van den Noortgate and Onghena (2008).

Because SSEDs result in small data sets, Ugille et al. (2013) suggested correcting the standardized effect sizes (i.e., b'_{ajk} in Equation 8.4) for small sample bias by multiplying the effect size by Hedges' bias correction factor (Hedges, 1981), which is approximately equal to 1-[3/(4*m*-1)], with *m* indicating the degrees of freedom. In the models discussed above, *m* equals the number of measurement occasions (*I*) minus the number of predictors (*p*) in the regression model minus 1 (i.e., $m = I - p - 1$):

$$b'^{C}_{ajk} = b'_{ajk}\left(1 - \frac{3}{4\,(I - p - 1) - 1}\right) \tag{8.5}$$

where b'^{C}_{ajk} represents the standardized, bias-corrected regression coefficient estimate for subject *j* from study *k*. Because the effect is multiplied by a constant, the variance of this effect size, of which the inverse is used as weight in a meta-analysis, should be corrected as follows:

$$\sigma^2_{b^c} = \sigma^2_b\left(1 - \frac{3}{4\,(I - p - 1) - 1}\right)^2 \tag{8.6}$$

8.6.3 Multilevel meta-analysis

Multilevel meta-analysis is one technique that can be used to combine effect sizes (standardized and bias-corrected regression coefficients in this study) measuring the same dependent variable such as the number of challenging behaviors during a specified time interval. The multilevel approach can be applied when synthesizing SSED studies in which at least some are characterized by more than one participant, because in that case, measurement occasions are nested within subjects and subjects in turn are nested within independent studies. The approach takes the hierarchical structure into account and the within-subject, between-subject and between-study variability in estimated intervention effect(s) can be

estimated. Such a multilevel analysis gives insight into the average intervention effects across subjects and studies, as well as variation in the effect between subjects and between studies and factors that explain this variation. Previous meta-analyses of single-subject studies ignored the different design types used in their primary studies and reduced all the designs into simple AB designs (e.g., Denis et al., 2011; Heyvaert et al., 2012; Heyvaert et al., 2014). This is unfortunate because in this way available information is not fully used. Below we re-analyze the data of Heyvaert et al. (2014) by means of three alternative models that take the specificities of the designs into account. To aggregate the data of multiple SSED types across subjects and across studies, the (standardized and bias-corrected) OLS regression coefficients can be combined either using a univariate meta-analysis for each kind of coefficient (i.e., b_1, b_2, b_3) or a multivariate meta-analysis of all types of coefficients together.

8.6.3.1 Univariate three-level meta-analysis

To combine the estimated effect sizes across subjects and across studies, three univariate multilevel meta-analyses can be performed, one for each kind of coefficient (i.e., ${b'}^{C}_{1jk}$, ${b'}^{C}_{2jk}$, and ${b'}^{C}_{3jk}$). This will be referred to as Model 1, although parameters are estimated for three separate univariate models– one for each coefficient type. At the first level of the multilevel model, these OLS estimated treatment effects equal an unknown population effect size, indicated by the β coefficients in Equation 8.7, plus a random deviation from this population parameter, indicated by the error terms:

Level 1:
$$
\begin{aligned}
{b'}^{C}_{1jk} &= \beta_{1jk} + e_{1jk} \quad \text{with} \quad e_{1jk} \sim N\left(0, \sigma^2_{e_{1jk}}\right) \\
{b'}^{C}_{2jk} &= \beta_{2jk} + e_{2jk} \quad \text{with} \quad e_{2jk} \sim N\left(0, \sigma^2_{e_{2jk}}\right) \\
{b'}^{C}_{3jk} &= \beta_{3jk} + e_{3jk} \quad \text{with} \quad e_{3jk} \sim N\left(0, \sigma^2_{e_{3jk}}\right)
\end{aligned}
\tag{8.7}
$$

The within-subject variability is set to the estimated variance from the ordinary least square regression analysis conducted for each subject.

The subject-specific population intervention effect, β_{1jk}, and differences between intervention effects, β_{2jk} for ATDs and β_{3jk} for ABAB phase designs, respectively, can vary from subject to subject and therefore we add a second level.

This results in the following:

Level 2:
$$\beta_{1jk} = \theta_{10k} + u_{1jk} \quad \text{with} \quad u_{1jk} \sim N\left(0, \sigma^2_{u_{1jk}}\right) \tag{8.8}$$
$$\beta_{2jk} = \theta_{20k} + u_{2jk} \quad \text{with} \quad u_{2jk} \sim N\left(0, \sigma^2_{u_{2jk}}\right)$$
$$\beta_{3jk} = \theta_{30k} + u_{3jk} \quad \text{with} \quad u_{3jk} \sim N\left(0, \sigma^2_{u_{3jk}}\right)$$

At the subject level the population regression coefficient for the intervention effect, β_{1jk}, equals an average intervention effect across subjects in the k^{th} study, θ_{10k}, and a subject-specific deviation, u_{1jk}, from the study-specific average intervention effect. Similar equations can be obtained for β_{2jk} and β_{3jk}.

At the study level, the study-specific effect sizes, θ_{10k}, θ_{20k}, and θ_{30k}, are allowed to vary across studies:

Level 3:
$$\theta_{10k} = \gamma_{100} + v_{10k} \quad \text{with} \quad v_{1jk} \sim N\left(0, \sigma^2_{v_{10k}}\right) \tag{8.9}$$
$$\theta_{20k} = \gamma_{200} + v_{20k} \quad \text{with} \quad v_{2jk} \sim N\left(0, \sigma^2_{v_{20k}}\right)$$
$$\theta_{30k} = \gamma_{300} + v_{30k} \quad \text{with} \quad v_{3jk} \sim N\left(0, \sigma^2_{v_{30k}}\right)$$

At level 2 and level 3, residual terms are assumed to be normally distributed. We are interested in γ_{100}, γ_{200}, and γ_{300} indicating the estimated effect sizes across subjects and across studies (the average intervention effect, the additional intervention effect of the second intervention for ATDs and the additional intervention effect during the second AB pair for the ABAB phase designs).

In a second model (Model 2), we are interested in assessing differences between design types in the average intervention effects. The analysis model is similar to the one presented using Equations 8.7 through 8.9. In both Model 1 and Model 2, three univariate three-level meta-analyses are performed, one for each effect size. The only difference between Model 1 and Model 2 is that the estimated intervention effect size, γ_{100}, will be separated into three estimated intervention effects, one for each design type. This can be accomplished by adding a dummy variable per design type in the univariate model estimating the average intervention effect: one for the AB/MBDs, one for the ATDs and one for the ABAB phase designs. The

dummy coefficient referring to the AB/MBDs is reduced to AB for simplicity. As a consequence, the first line of Equation 8.7 becomes:

$$b'^{C}_{1jk} = \beta_{1jk,AB}AB_{jk} + \beta_{1jk,ATD}ATD_{jk} + \beta_{1jk,ABAB}ABAB_{jk} + e_{1jk} \,. \qquad (8.10)$$

8.6.3.2 Multivariate three-level meta-analysis

The multivariate model is of importance when the researcher wants to estimate multiple effect sizes simultaneously (as opposed to conducting separate analysis per type of effect size as presented in the univariate section), taking into account that effect sizes within a subject are possibly correlated, as is the case for the ABABs (between γ_{100} and γ_{300}) and the ATDs (between γ_{100} and γ_{200}). Another major advantage is that, in addition to variance estimates, the multivariate model allows estimation of the covariance between effect sizes at the subject and study level. For instance, the covariance between the average intervention effect estimate and the additional intervention effect estimate for the ATDs and the ABAB phase designs at the subject level and study level can be estimated (Van den Noortgate & Onghena, 2003b). The multivariate three-level meta-analysis might be preferred because of its ability to provide more information compared to the univariate three-level model approach as well as its handling of the potentially non-zero covariances between effect sizes.

The estimation approach is a multivariate extension of the univariate approach of Model 1. Kalaian and Raudenbush (1996) describe how such a multivariate analysis can be executed, using estimates of the sampling variances and covariances of the observed effect sizes (here, estimates obtained from the OLS regression analyses). The covariance between effect sizes at the second and third level will be estimated. This multivariate model is the third analysis model (Model 3) we discuss in the empirical illustration.

8.7 Empirical Illustration

Recently, Heyvaert, et al. (2014) conducted a meta-analysis of 59 SSEDs looking at restraint interventions for challenging behavior among persons with intellectual disabilities. Heyvaert et al. (2014) retrieved raw data graphically presented in the primary studies using the statistical software program UnGraph, recommended by Shadish, et al. (2009). In the study of Heyvaert et al. (2014), the previously discussed SSEDs are included (AB, ABAB phase designs, MBDs, and ATDs). However, they simplified their dataset by reducing all types of SSEDs to simple AB phase designs. In this study, we re-analyze 55 of the 59 studies (i.e., 4 studies were not re-coded because it was not clear which design type was used in the

primary study), and recoded them (taking multiple intervention phases and multiple interventions within a phase into account). We found that within some studies, multiple types of SSEDs were used. For instance the first subject can be characterized by an ABAB phase design and another subject by an ATD. A total of 98 subjects were recoded and a median of 1 subject per study was found with a minimum of 1 and a maximum of 6 subjects per study. The median number of measurements within a subject was 51 with a minimum of 9 and a maximum of 237. Over one half of the subjects (52.82%) were characterized by an ABAB reversal design, 22.45% of the subjects were associated with simple AB phase designs, 18.37% of the subjects are MBDs, and 6.36% are ATDs. For the ATDs, the two introduced treatments per subject are variants of each other.

Table 8.1 gives an overview of the parameter estimates and standard errors when using the different analysis models of interest. Model 1 and 2 are composed of three univariate three-level meta-analyses. In Model 1, an average intervention effect is estimated across the SSED types (γ_{100}), and an additional intervention effect for the second AB pair was estimated for the ABAB phase designs (γ_{300}) and an additional intervention effect for the second treatment was estimated from the ATDs (γ_{200}). In Model 2, predictors indicating the design type are included. This results in an average intervention effect estimate per design type ($\gamma_{100,AB}$, $\gamma_{100,ABAB}$, and $\gamma_{100,ATD}$). Again, an additional intervention effect for the ATDs (γ_{200}) and the ABAB phase designs (γ_{300}) was estimated. In the third analysis model (Model 3), the dependency between effect sizes was modeled by using the multivariate three-level model.

The average intervention effect estimate across SSED types was estimated in Model 1 and Model 3 and was found to be statistically significant: -3.26, $t(42) = -7.23$, $p < .01$, and -3.12, $t(41) = -15.99$, $p < .01$, respectively. This indicates that in general, restraint interventions are effective in reducing challenging behavior among persons with intellectual disabilities. The average intervention effect estimates per SSED type (i.e., Model 2) was found to be statistically significant for the AB/MBDs and the ABAB phase designs, but insignificant for the ATDs (see $\gamma_{100,AB}$, $\gamma_{100,ABAB}$, and $\gamma_{100,ATD}$ in Table 8.1). The three models allow estimating whether the intervention effect in the second AB pair (for ABAB phase designs) differs significantly from the intervention effect in the first AB pair, which is not the case in the univariate models. In contrast to the statistically non-significant effect found in Model 1 and Model 2, the multivariate model indicates that the additional intervention effect for the ABAB phase designs is statistically significant; γ_{300}= -1.10, $t(12) = -5.62$, $p < .01$. The additional effect of the second intervention (for ATDs) is not statistically significant.

Heyvaert et al. (2014) also estimated the average intervention effect across subjects and across studies, but simplified all included SSEDs to simple AB phase designs. They found an average significant intervention effect estimate of -3.16, which is comparable with the one obtained by Model 1 and Model 3. However, Heyvaert et al. (2014) did not provide separate effect size estimates for each SSED type nor any information concerning additional intervention effect estimates.

No statistically significant variance is found at the study level for the univariate models (i.e., Model 1 and Model 2), whereas significant variance estimates are found for the multivariate model (i.e., Model 3). At the subject level, all variances are statistically significant except for the additional intervention effect estimate for the ATDs. For the univariate models, the between-subject variance of the average intervention effect estimate is more than seven times larger than the between-subject variance of the additional intervention effect estimates, which is not the case for the multivariate model. In addition to variance estimates, the multivariate model allows estimation of covariance between effect sizes. We found a statistically significant positive covariance between the average intervention effect estimate and the additional intervention effect estimates both at the subject and study level. For instance, the covariance between the average intervention effect estimate and the additional intervention effect estimate for the ATDs at the subject level equals 1.03, $Z = 4.95$, $p < .01$. In order to interpret this value, we can calculate the correlation. The correlation between the two effect sizes equals the covariance of these effect sizes divided by the product of the standard deviation of both effect sizes. In this example, the correlation between the average intervention effect estimate and the additional intervention effect estimate for the ATDs at the subject level equals 0.65 $\left[\text{i.e.}, 1.03 \,/\, (\sqrt{1.48}.* \sqrt{1.68})\right]$. This means that a large average intervention effect estimate goes together with a large additional intervention effect estimate. Note that the estimated variances in the multivariate model are smaller compared to the estimated variances using the univariate models, but still statistically significant because of the smaller estimated standard errors. The estimated variance components of the study of Heyvaert et al. (2014) are comparable to the ones obtained by Model 1 and 2. In their study, the between-study variance was not statistically significant (= 3.49), but the between-subject variance was significant (= 12.21). The study of Heyvaert et al. (2014) could not provide any information concerning the variance of additional intervention effects or the covariance between effect sizes.

Table 8.1

Parameter and Standard Error Estimates Resulting from Estimation of the Three-Level Meta-Analysis of Model 1 – Model 3

Parameter	Parameter estimate (*SE*)		
	Model 1	Model 2	Model 3
Fixed coefficient			
Average intervention effect			
across SSED types, γ_{100}	-3.26* (0.45)		-3.12 *(0.20)
per design type			
$\gamma_{100,AB}$		-3.91* (0.70)	
$\gamma_{100,ABAB}$		-2.72* (0.58)	
$\gamma_{100,ATD}$		-2.00 (2.27)	
Additional intervention effect			
Second intervention effect in ATD, γ_{200}	-1.83 (0.90)	-1.83 (0.90)	-0.96 (0.42)
second AB pair for ABAB phase designs, γ_{300}	-0.12 (0.07)	-0.12 (0.07)	-1.10* (0.20)
(co)variance component			
Between-study variance			
Average intervention effect, $\sigma^2_{v_{10k}}$	3.00 (2.18)	3.30 (2.22)	1.07* (0.15)
Average additional intervention effect			
Second intervention effect in ATD, $\sigma^2_{v_{20k}}$	0.50 (2.96)	0.50 (2.96)	1.00 (0)
Second AB pair for ABAB phase designs, $\sigma^2_{v_{30k}}$	0.0 (-)	0.0 (-)	1.14* (0.16)
Covariance intervention effect second intervention in ATD			0.99 (0)
Covariance intervention effect second AB pair			1.11* (0.16)
Between-case variance			
Average intervention effect, $\sigma^2_{u_{1jk}}$	13.12* (2.52)	12.85* (2.49)	1.48 * (0.10)
Average additional intervention effect			
Second intervention effect in ATD, $\sigma^2_{u_{2jk}}$	1.83 (1.98)	1.83 (1.98)	1.68 (0.73)
Second AB pair for ABAB phase designs, $\sigma^2_{u_{3jk}}$	0.17* (0.05)	0.17* (0.06)	1.42* (0.11)
Covariance intervention effect second intervention in ATD			1.03* (0.21)
Covariance intervention effect second AB pair			1.22*(0.09)

Note. Standard errors are in parentheses.
* $p < .05$.

8.8 Discussion

The purpose of this study was to illustrate the multilevel meta-analysis of SSEDs of different types because current methodological work on the synthesis of SSEDs often focuses on combining simple AB phase designs and MBDs or simplify more complex SSEDs to simple AB phase designs. We focused on the average intervention effect estimate and we proposed three different analysis models to accomplish this and compared the results to the study of Heyvaert et al. (2014). We found a significant average intervention effect across SSED types (Model 1 and Model 3). This matches the results of the study of Heyvaert et al. (2014). When the average intervention effect is estimated per design type (Model 2), an insignificant effect is found for the ATDs, which are represented in only 6% of the studies and thus their effect is estimated less precisely than the effect for AB or ABAB designs. This is a major limitation of Model 2. The multivariate model (i.e., Model 3) gives more information in comparison to the univariate one because the dependency between different kinds of effect sizes is handled and estimated. For instance, in this study we found that the magnitude of the average intervention effect was strongly positively correlated with the additional

intervention's effect. A major advantage of the multivariate model is that the average effect sizes can be estimated simultaneously, whereas in model 1 and 2, three univariate multilevel meta-analyses have to be performed.

This study has some limitations. The focus of the current study was on synthesizing results from SSED studies involving different design types and investigating average effects. However, it is possible to also use the multilevel model demonstrated here to provide subject-specific estimates. To extend the multilevel meta-analysis of SSEDs to incorporate multiple design types we chose to start with a relatively simple statistical model. We assumed that the within-subject variance was homogenous over phases, but in some studies the variance may change with intervention. We also assumed that the within-participant errors were independent, however, errors may be autocorrelated (Beretvas & Chung, 2008; Ferron et al., 2009; McKnight, McKean, & Huitema, 2000). We assumed that the errors at the three-levels were (multivariate) normally distributed, which may not be the case with some outcomes (e.g., counts of behaviors). We also assumed no trend in baseline and treatment phases. However, in some studies a linear or non-linear trend might be more realistic. More complex regression and multilevel models have been proposed to handle these various data complexities, but we chose to start with a relatively simple statistical model. Now that we have developed a multilevel meta-analytic method for synthesizing results across design types, future research could extend the approach to more complex models that can accommodate these additional data complexities. Making the methods for synthesizing SSED studies more inclusive of varying SSED design types and varying data structures will lead to more precise and accurate estimates of average treatment effects, as well as more precise and accurate estimates of the variance of treatment effects across participants and studies.

PART 3| DISCUSSION, CONCLUSION AND FUTURE RESEARCH

Chapter 9|
General Discussion

9.1 Introduction

SSEDs have been used to evaluate the effect of a treatment on the outcome score of a dependent variable for many years (Busse et al., 1995; Chorpita et al., 1996; Kratochwill & Levin, 1992). However, only during the last decade, SSEDs have been acknowledged as a means to establish an evidence base for examining the effectiveness of treatments (Kratochwill et al., 2010). As a consequence, it is not surprising that especially during recent years the number of published SSEDs has been increasing at an astonishing rate in accordance with the need of appropriate techniques to analyze SSED data. In order to support an evidence base and make externally valid conclusions to inform research, practice and policy it is not only necessary to analyze SSED data but also to meta-analyze (i.e., summarize, synthesize) a set of similar focused SSEDs. The literature indicates that there has been made efforts to meta-analyze SSEDs as the number of published meta-analysis of SSED studies is increasing, especially during the last seven years (e.g., Social Science Citation Index within the Web of Sciences). However, in contrast to meta-analysis of group-comparison studies, there is still a lack of consensus regarding methods to quantitatively summarize SSED results. Over the past twenty years, suggestions have been made about how to meta-analyze these SSEDs (Center et al., 1985-1986; Hedges et al., 2012; Maggin et al., 2011; Mastropieri & Scruggs, 1985; Parker et al., 2011; Shadish et al., 2013; Shadish, et al., 2012; Swanson & Sachse-lee, 2000), but it is fraught with a lot of controversy (Allison & Gorman, 1993; Kratochwill et al., 2010; Shadish & Rindskopf, 2007). To combine SSED data within and across studies, a promising approach recently suggested and recommended is multilevel modeling (Nugent, 1996; Rindskopf & Ferron, in press; Shadish & Rindskopf, 2007; Shadish et al., 2013; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). The multilevel modeling method is a very flexible approach given its ability to model complexities such as autocorrelation, predictors at the different levels (such as age, gender, school type, study quality), heterogeneous within-subject, between-subject and between-study covariance and allows estimating average treatment effects in addition to subject-specific and study-specific treatment effects (Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003b). By conducting a multilevel analysis, important research questions can be addressed (which cannot be answered by single-level analysis of SSED data) such as: (1) What is the magnitude of the average treatment effect across subjects and across studies? (2) What is the magnitude and direction of the subject-specific intervention effect? (3) How much does the treatment effect vary within

subjects, across subjects and/or across studies? and (4) Does a (subject and/or study level) predictor influence the treatment's effect?

This dissertation is dedicated to the three-level modeling of SSEDs as little is known about its potentials, modeling options, performance and power. The goal of this dissertation is twofold. The first part **(***Part 1***)** deals with the examination of the performance of the three-level model to summarize SSED data across subjects and across studies and is especially interesting for methodologists, meta-analysts, and single-case research synthesists (*Chapters 2, 3, 4, and 5*). The second part (*Part 2)* focuses on modeling options and practical applications of the three-level model and is intended for an applied audience (*Chapters 6, 7, and 8*). In the following sections, an overview is given of the main research findings from the different studies and their strengths and limitations are discussed (*Chapter 9*). *Chapter 9* is of primordial importance as it handles implications for SSED researchers, SSED meta-analysts, and methodologists. Research without implications would be meaningless. We end this chapter with a brief summary and global conclusion. *Chapter 10* is the most important chapter of the dissertation as it yields suggestions to continue research in the area of multilevel analysis and single-subjects.

9.2 Research Overview: Summary of the Main Findings

9.2.1 Part 1

The first part is methodological and presents the results of four computer-intensive Monte Carlo simulation studies.

In *Chapter 2*, the first Monte Carlo simulation study (of which three will follow) was conducted in order to evaluate whether the basic three-level model (including changes in level, changes in slopes and homogeneous variances at the three-levels) is appropriate to synthesize SSED data across subjects and across studies. The simulation study shows that the three-level approach results in unbiased estimates of the immediate treatment effect and the treatment effect on the slope. In order to have reasonable power ($\geq$ .80) for testing the treatment effects, a homogeneous set of at least 30 studies should be included. If this condition is fulfilled, the number of measurements and subjects is of less importance. From the estimates of the variance components, we deduce that there is bias when estimating the between-case variance of the immediate treatment effect and the effect on the slope (mean bias equals 2% with a minimum of 0% and a maximum of 10%). These biased variance estimates are consistent with previous empirical research about the three-level analysis of

SSED data (Owens & Ferron, 2012) and previous research from a broader methodological domain, for instance growth curve models (Kwok et al., 2007; Murphy & Pituch, 2009). The bias is even larger when estimating the between-study variance of the immediate treatment effect and the effect on the time trend (mean relative bias equals 17% with a minimum of 0% and a maximum of 49%). In these contexts the inclusion of 30 or more studies is recommended, and even then researchers should anticipate some bias.

In *Chapter 3***,** the focus of interest is on one specific method, originally suggested by Van den Noortgate and Onghena (2008), to standardize raw single-subject data. They proposed performing an ordinary least squares regression for each case separately, and dividing the individual scores by the estimated residual within-case standard deviation. The issue of standardization is timely, as there is an increasing interest in combining SSED data across a variety of different studies, and as dependent variables in a set of SSED studies are not always measured the same way and on the same scale. Therefore standardization is needed to allow immediate comparison and fair interpretations of scores on challenging behavior across different studies. The three-level synthesis of standardized single-case data is found to be appropriate for the estimation of the treatment effects, especially when many studies (30 or more) and a lot of measurement occasions within cases (20 or more) are included, and when the studies are rather homogeneous (with a small between-study variance). The estimates of the variance components are less accurate. There is a significant mean bias for the estimate of the between-study variance (24%) and the between-case variance (71%) for the immediate treatment effect (similar results are obtained for the estimated between-study and between-case variance for the treatment effect on the slope). The maximum bias for the estimation of the between-study variance is 35%: while it is 219% for the estimation of the between-subject variance. Especially the estimate of the between-subject variance is biased, except when at least 40 measurement occasions within a case are included. The estimation of the between-study variance is more accurate, but the bias remains substantial if 30 studies with only 10 measurements within a subject are included. The bias is especially larger when the between-study variance is large and the within-study variance is small. For both estimated variance components, the number of measurements has a large effect on the bias.

In *Chapter 4*, we proposed a possible way to handle a threat towards internal validity, commonly encountered in SSED, namely external event effects. In multiple-baseline designs, external effects can become apparent if they simultaneously have an effect on the outcome score(s) of the cases within a study. This study presents a method to adjust the three-level

model for external events and evaluates the appropriateness of the modified model. The results of the simulation study show that if the external event influences subsequent scores for all the cases within a study, the three-level approach for uncorrected effect sizes is not recommended because the estimates of both treatment effects (i.e., immediate effect on level and effect on time trend) are biased. The mean squared error, standard error, and coverage proportion are better estimated when using the modified model, which includes moment effects. The difference between the corrected and uncorrected effect sizes is largest when there are a small number of studies (10) and measurement occasions (15), so in this context we advise using the adjusted model. Moreover the adjusted model results in less biased variance estimates. As was found here, even when an external event effect is small, a failure to correct for it can lead to biased effect sizes. For instance, the maximum relative bias equals 313% for the estimated between-study variance of the immediate treatment effect for the uncorrected effect sizes, while it is 55% for the corrected effect sizes. Thus, single-subject data analysts are encouraged to consider use of the three-level model that corrects for external event effects when synthesizing results of multiple-baseline design data.

In the last chapter of the first part**,** *Chapter 5***,** the robustness of the multilevel model against misspecification of the covariance matrix at the second and third level of the three-level model is investigated. The results confirm previous research and indicate that the treatment effect estimates across subjects and across studies are unbiased. However, when covariance is generated, the mean squared error is large, but is reduced by increasing the number of studies and subjects and reducing the between-subject variance. The median relative standard error biases is also substantial when covariance is generated and only slightly larger when covariance is generated and ignored in the analysis. As a consequence, the coverage proportion of the 95% confidence intervals is too small. This indicates that the treatment effect estimates across cases and across studies is relatively robust against misspecification of the covariance matrix. However, this is not the case for the estimates of the variance components (i.e., between-case and between-study variance). When covariance is present in the generated data but ignored in the analysis, the between-study variance and between-subject variance has large bias values. Thus, when researchers are interested in the estimate of the variance components, modeling covariance at the second and third level is recommended.

9.2.2 Part 2

The second part of this dissertation is the applied part and contains three applications of the multilevel modeling method.

In *Chapter 6*, representing the first applied chapter, we present and extend the piecewise regression analysis (Center et al., 1985-1986) to analyze SSED data. The regression-based approach is a flexible technique to analyze SSED data retrospectively, in addition to visual analysis during data collection. The purpose of the regression analysis is to quantify the SSED data results resulting in an effect size estimate which can be used to compare SSED results across studies, enhance the communication between applied SSED researchers, and can be used in meta-analysis to synthesize a large body of research. But, one question is how to specify predictors in a regression model in order to account for the specifics of the design and estimate the effect size of interest. In this study, we go back to basics and discuss in large detail what the design matrix looks like for the three most popular design matrices (i.e., multiple-baseline designs, reversal designs, and alternating treatment designs) to estimate the regression coefficients of interest. A graphical presentation of the regression coefficients is displayed and empirical illustrations are given.

In *Chapter 7*, the extension of a single-level analysis to a three-level analysis of SSED data is discussed in detail. The enormous flexibility of the multilevel models is also a major drawback, because there are a variety of different modeling options and it is not always obvious in which conditions which to choose. Therefore in this article, a variety of different modeling options are discussed and illustrated using real datasets. We investigate to what extent the estimated treatment effect is dependent on the modeling specifications and the underlying assumptions. By considering a range of plausible models and assumptions, researchers can determine the degree to which the effect estimates and conclusions are sensitive to the specific assumptions made. If the same conclusions are reached across a range of plausible assumptions, confidence in the conclusions can be enhanced. We advise researchers not to focus on one model but to conduct multiple plausible multilevel analyses and investigate whether the results depend on the modeling options.

In *Chapter 8*, the multilevel modeling framework is extended by giving different modeling options to combine several different SSED types (i.e., multiple-baseline design, ABAB reversal designs, and alternating treatment designs). This is timely, as in previous research all SSEDs are commonly reduced to simple AB phase designs, ignoring the complex data structures, which possibly resulted in biased treatment effect estimates. We illustrate the

different models using a re-analysis of a meta-analysis of single-subject experimental designs (Heyvaert et al., 2014). The intervention effect estimates using univariate three-level models differ from those obtained using a multivariate three-level model that takes the dependence between effect sizes into account. Because different results are obtained, and because the multivariate model has multiple advantages – including more information and smaller standard errors – we recommend researchers to use the multivariate multilevel model to meta-analyze studies that utilize different SSED types.

9.3 Strengths and Limitations of this Dissertation

9.3.1 Strengths of this dissertation

The first major strength of this dissertation lies in validating a promising flexible technique to summarize a large body of literature, namely multilevel modeling. The further development of such a technique is timely as there is a growing interest in summarizing SSED studies as a means to establish an evidence base for the effectiveness of treatment effects. The number of published SSED studies has been increasing at a spectacular rate during the last decade, which makes them suitable for quantitative syntheses. Moreover, it would be a waste of research investments to neglect this wealth of information that is already in the literature. The multilevel model can be used to perform SSED data synthesis and takes the hierarchical structure of the data into account. It is surprising that previous research did not use the multilevel model to synthesize SSED data, but rather focused on subject-specific analysis. A reason for this might be that the multilevel model methodology was not well understood and was not further developed; problems that are hopefully solved, at least partially, after reading this dissertation.

A second major strength of this dissertation is that we focus on a heterogeneous audience, namely methodologists, meta-analysts, and applied single-case researchers. The articles included in the dissertation are published in both methodological and applied international peer-reviewed journals, and therefore we tried to make the multilevel modeling option widely-spread known. In the first part of this dissertation, methodologists are challenged by the extensive and computer-intensive simulation studies conducted using the infrastructure of the Flemish Supercomputer Center. Given the urgency, we challenge methodologists to investigate further developments and complexities resulted from the simulation studies. A lot of coding expertise and statistical knowledge is needed. Because new methodologies are meaningless without applications, we elaborated a second part in which

applied single-case researchers and meta-analysts are taken on a journey through the multilevel modeling process using a step-by-step approach and making the multilevel model gradually more complex. Different modeling options are presented and illustrated using real data examples. We explain in large detail why research synthesis and multilevel modeling is needed.

A third strength of this dissertation is that the included conditions in the simulation studies represent realistic conditions, as they are based on thorough re-analysis of published meta-analyses of SSEDs in the field of educational research. During my first months as a *PhD*-student, raw data were retrieved from these meta-analyses of SSEDs using statistical software, which resulted in a dataset allowing identification of realistic conditions and parameter values. We also consulted recently published articles such as the Shadish and Sullivan (2011) paper and the *WWC* technical documentation (Kratochwill et al., 2010) giving characteristics of SSEDs. We did not only formulate conditions to be included in the basic three-level model, but we also formulated complexities such as external event effects, misspecification of the covariance matrix and the issue of standardization. All these extensions are discussed in detail and are validated through extensive Monte Carlo simulation studies. In the first two simulation studies (*Chapter 2* and *Chapter 3*), we included 2,000 iterations of each condition, which reduces the likelihood that the results are obtained by chance. For the other two simulation studies, we only included 500 (*Chapter 5*) or 400 (*Chapter 4*) replications. Because of the model complexity, and time and money issues, we had to find a balance between the number of replications and the included conditions and parameter values in these last two simulation studies. Although we did not empirically validate all possible extensions to the three-level model, a major strength of this dissertation is that we discussed a variety of different and plausible modeling options in the applied part (i.e., we chose to model autocorrelations, heterogeneous within-case variance, predictors at the different levels, trends during the treatment phase, and different types of SSEDs designs such as multiple-baseline designs, reversal designs and alternating treatment designs). By considering a range of plausible models and assumptions, researchers can determine the degree to which the effect estimates and conclusions are sensitive to the specific assumptions made. If the same conclusions are reached across a range of plausible assumptions, confidence in the conclusions can be enhanced.

Previous research aiming to summarize SSED data only focused on fixed effect estimates ignoring the meaningful information (co)variance components estimates can provide. An additional strength of this dissertation is that the (co)variances are modeled and

estimated, in addition to the fixed effect estimates, and the influence of the covariance specification on the treatment effect estimates and (co)variance components estimates is evaluated. For instance, the between-study variance indicates to what extent the estimated treatment effects of individual studies deviate from an average treatment effect estimate. We found that homogeneous studies have a beneficial effect on the treatment effect estimates. Also, the covariance between regression coefficients at the different levels should not be ignored, as it can give interesting information. For instance, a non-zero covariance between residuals at level 2 seems reasonable, as due to a ceiling effect, the estimated treatment effect is expected to be smaller for cases with an already high estimated baseline level.

The traditional estimation procedure used within the multilevel modeling framework is the restricted maximum likelihood estimation (REML). A strength is that the REML estimation of multilevel models is implemented by default in a variety of commercial software programs such as SAS, HLM, MLwiN, SPSS, and Stata, and is even available in the freeware, R. This enhances the use of multilevel modeling. Although we only elaborated on how to conduct multilevel modeling using SAS codes, a strength of this manuscript is that a description of the codes is provided, enabling the researcher to make the translation to the preferred statistical software program.

A last major strength of this dissertation is that it is embedded in an international context. This dissertation is the result of a collaboration between the Methodology of Educational Science Research Center of KU Leuven, Quantitative Methods, University of Texas, and Educational Measurement and Research, University of South Florida. As a consequence, the research team is composed of researchers with a different background: experts in multilevel modeling, SSEDs, meta-analysis, randomization tests, and applied SSED research. This mix of research expertise enabled the development of methodological innovations that are relevant for everyday SSED practice. Also because of the international interest, the use of the multilevel modeling technique can be spread widely.

9.3.2 Limitations of this dissertation

A first problem is that the multilevel model assumes that studies and subjects are randomly sampled from a population of studies and subjects respectively. However in reality and practice, subjects are not randomly sampled, but are chosen on purpose because the researcher is interested in that particular subject (e.g., subjects with special needs). As a consequence, researchers interested in summarizing SSED studies must be careful in generalizing SSED results to a broader population of subjects. It is important to define the population, as conclusions are exclusively restricted to the population of subjects from which the included subjects in the multilevel analysis could be regarded as a random sample. Note that this is not only a limitation of SSED studies. Also in group-comparison designs, participants can be purposely sampled.

Although the conditions included in the simulation studies are chosen to be representative for published SSED studies in the domain of education, as we conducted re-analyses of published SSEDs in this research area, not all conditions could be included in the Monte Carlo simulation studies. As a consequence, we evaluated the appropriateness of inferences made from a three-level single-subject model in specific conditions and caution is warranted with generalization of the research findings to conditions that were not simulated. In addition, in this dissertation we focused on balanced data. We chose to keep the number of measurements within a study constant for all subjects within the same study. Of course it is possible that different participants of the same study have different series lengths. Also homogeneous within-subject, between-subject and between-study variance is assumed whereas this might not be the case.

The number of replications (i.e., datasets) per condition in this simulation study is either 2,000, 500 or 400 depending on the model complexity and the number of included conditions (i.e., if models are rather basic, 2,000 replications were feasible, otherwise 500 or 400). Of course, as more replications emerge, one can be more confident in the research findings. If money and time issues are not a matter of concern, we advise to include 2,000 replications (or even more). However, the chance is small that the results are obtained by coincidence if 500 or 400 replications are included.

Another issue in the context of multilevel analysis of SSEDs is that small sample sizes are an inherent characteristic of SSEDs, which limits the achievable complexity of the three-level model. As models are getting more complex (as in the fourth simulation study where covariance at the second and third level is modelled), and represent more realistic models that

fit the data better, the estimation might fail due to small sample sizes. This might have been the reason why in the fourth simulation study, ignoring existing covariance did not have consequences on the fixed effects and variance components estimates.

Although our research team consists of several researchers with a complementary background, a person with extended algebraic knowledge is missing. It might be possible to mathematically derive large-sample approximations of the estimated standard errors of the treatment effects for balanced situations. Because of the small sample sizes in the context of multilevel modeling of SSED data, asymptotic assumptions are violated, upon which the algebraic derivations would be based. However, algebraic derivations could have provided guidance when setting up the simulation studies or could have helped the interpretation of the simulation results. Thus, we exclusively rely on simulation studies to empirically examine the three-level modeling technique.

In this dissertation, we only discovered the top of the iceberg and validated the three-level model with relatively few extensions (i.e., modeling trends, standardization, external event effects, and covariance at the second and third level). Other extensions are plausible and some of them were proposed in the applied part of the dissertation, such as autocorrelation, heterogeneous variance, non-linear trajectories, count outcomes, etc. However, little is known about the performance of the three-level model if extensions and combinations of extensions are included. So far, the multilevel model has been promising, but caution is needed when stating that the multilevel model is appropriate to summarize SSED data because we did not focus on all possible extensions and combinations of extensions.

We simulated data assuming that the errors at the different levels are independent, identical, and (multivariate) normally distributed. First, there are concerns regarding the assumption that the errors in the statistical model are independent. When repeated observations are made on the same subject, it is plausible that the errors of the measurement associated with a score at one data point may be predictive of errors at other points in the series that follow. To simplify the simulation model, we did not account for a possible dependence between different regression coefficients, which can be accounted for in a multilevel analysis by estimating the covariances at the various levels. It is likely that the distribution of the errors are case-and study-specific. Lastly, because of the small sample sizes included (especially at the second level of the multilevel model), it is hard to examine the normality assumption.

A next limitation lies in the fact that we only focused on the traditional estimation procedure in multilevel modeling for estimating fixed effects and variance components,

namely the restricted maximum likelihood. We did not include alternative estimation procedures such as bootstrapping and Bayesian estimation procedures (Rindskopf, 2013; Van den Noortgate & Onghena, 2005). These procedures are promising as they rely on less restrictive assumptions but are not yet investigated in the context of three-level modeling of SSED data. Alternative estimation techniques might solve the biased variance component estimates. Despite the valuable information these variance components provide, we did not yet focus on possible solutions to deal with this issue.

In this dissertation, we only focused on educational and social sciences data. However, the number of published SSEDs is increasing in a variety of different research fields, including for instance the biomedical world. We only discussed educational and social sciences data because they are different in nature from biomedical data. However, despite the particular characteristics of educational data, a lot could have been learned from other research fields. It is unfortunate that we did not look beyond our own research field and broaden our view.

Another limitation is that we focused on the multilevel modeling of raw SSED data. This is only possible if raw data can be retrieved from the original SSED studies. Usually this will be the case as there is a tradition in SSED research to present the data graphically. However, retrieving raw data from single-case studies using statistical software tools (which is a point and click procedure) entails a large workload. Also, it might be the case that the data are not graphically presented and that the author of the SSED study is unable to provide the raw data. In that case, combining effect sizes instead of raw data is a solution as discussed by Ugille et al. (2012). However, it might be the case that the effect sizes are not reported in the original studies and additional calculations have to be conducted.

9.4 Implications of this Dissertation

9.4.1 Implications for research synthesists

The results of this dissertation are promising and encouraging for researchers interested in estimating fixed effect (i.e., average immediate treatment effect and treatment effect on slope) across subjects and across studies. Valid and reliable average treatment effect estimates are obtained at least if the underlying assumptions are met and if the model is correctly specified. The results of these syntheses establish a means of evaluating treatment effect estimates and contribute to evidence based practice. Valuable information is obtained in order to improve research and everyday practice and important policy decisions can be made based on the results of literature synthesis. However, we advise research meta-analysts to increase the number of primary studies included in the multilevel analysis whenever possible as greater precision and accuracy in effect size estimates can be obtained. While single-case meta-analysts are constrained by the availability of primary studies, they could adjust their methods for searching (e.g., expanding their search terms) whenever possible, but are limited by what the field has generated.

In contrast to the average fixed effect estimates, caution should be paid when interpreting the variation in treatment effects between subjects and between studies. Even assuming the model is correctly specified, the variance components at all levels are biased. Furthermore, it is not always obvious which specific SSED data characteristics and synthesis characteristics best fit the data and have to be modeled. Therefore, we advise the research synthesists to conduct a sensitivity analysis and evaluate to what extent the average treatment effect estimates depend on specific modeling options. If the same conclusions are reached across a variety of different multilevel modeling options, the researcher can be more confident in its research findings and more reliable estimates are obtained.

9.4.2 Implications for applied single-case researchers

The study shows that the average treatment effects are generally well estimated if the between study-variability is small and if a minimum of 30 studies are involved. The number of measurements and cases is of less importance. Therefore, besides the importance of systematically varying characteristics of studies in order to investigate moderator effects, it might be advantageous to replicate previous studies, resulting in homogeneous study results. Of course the methodology and instruments have to remain appropriate for the subject in a certain context. Also, single-subject researchers should pay attention to baseline variability or stability in an effort to decrease variability at level one. This might partly solve

standardization problems. Also, as the baseline trajectory is used as a means to estimate the treatment effect, a more justified treatment effect estimate is obtained. Another cause of variability is measurement error. Finding a way to eliminate measurement error might decrease overall variability. Therefore, we encourage single-case researchers to measure a dependent variable consistently at the same time of day, at the same setting and for the same amount of time across subjects and even across studies investigating the same underlying treatments. We also advise single-case researchers to pay attention to treatment fidelity (Kazdin, 2011), because this can result in a decrease in between-subject variability, and as a result in less variability in the average treatment effect estimate. For example, if a treatment was administered exactly like it was intended to be administered, the associated treatment effect would be different than a treatment effect associated with a treatment administered differently than intended.

A requirement for obtaining more accurate estimated treatment effects over subjects and over studies if standardized SSED data are used, is to include at least 20 measurement occasions per subject. As standardization is desirable if SSED studies are combined, we encourage single-subjects researchers to observe and measure their subjects at least 20 times. A final recommendation to single-case researchers is to consider previous single-subject studies. Specifically, if single-case researchers from similar areas of interest (e.g., reading, math) measure their dependent variables the same across studies, then single-case meta-analysts would have a larger number of primary studies to include in their research synthesis and could feel more confident in their interpretation of average treatment effect estimates.

9.4.3 Implications for methodologists

We encourage methodologists studying the use of multilevel modeling to summarize single-case data to conduct further research on modeling count outcomes, non-linear trajectories, etc. Furthermore, violations of assumptions (e.g., non-normality of the level-1, level-2, or level-3 errors, heteroscedasticity of errors at all levels) and various level-1 error models (e.g., high order autoregressive or moving average models) need to be investigated in the future. Investigation of these more complex models would allow for a better understanding of the applicability of the models under a variety of conditions. Future research on other approaches to estimate variance components would also be of interest. The results of this dissertation have indicated that the variance components at all levels are biased. Therefore, it would be interesting to investigate alternative methods for estimating variance

such as the Bayesian approach. More details about these suggestions and other suggestions for further research are given in *Chapter 10*.

9.5 Global Conclusion

From the first part of this dissertation, we conclude that the basic three-level multilevel model is found to be appropriate to synthesize raw unstandardized single-case experimental data across cases and across studies, at least if the researcher is interested in the fixed effect estimates (i.e., immediate treatment effect and treatment effect on slope) and if the underlying assumptions are not violated (independent, identically and normally distributed errors). The variance components estimates (i.e., the between-case and between-study) can be questioned and are biased, especially when estimating the between-study variance of the immediate treatment effect and the effect on the time trend. The bias can be reduced by including at least 30 studies, but even then researchers should anticipate some bias. When the basic three-level model is extended by modeling standardized raw data instead of unstandardized raw data, the variance components estimates become even more biased. This dissertation has shown that the multilevel modeling approach can easily be extended, by for instance taking external event effect into account, a major threat towards internal validity in SSED research. Moreover, the multilevel model is relatively robust against misspecification of covariance components at the second and third level of the multilevel model. This first part should be extended in further research, by extending the basic three-level model by including combinations of commonly encountered complexities, such as standardization in combination with external event effect, autocorrelation, non-linearity, etc.

From the second part of this dissertation, we conclude that the multilevel modeling framework has a lot of applications and a lot of potential. Both the two- and three-level model can easily be extended by taking for instance count data, autocorrelation, non-linearity, heterogeneous within-phase variance, etc. into account. However, caution should be paid when specifying the design matrix, as the interpretation of the parameters of interest is dependent on this. There are also different ways to combine different types of SSEDs using multilevel models. So its enormous flexibility is one of the major advantages of the multilevel model, but at the same time one of its major pitfalls.

Chapter 10|The future of Multilevel Modeling to Synthesize Single-Subject Experimental Design Data?

Suggestions for Further Research

Doing research is a never-ending process and does not simply end with a global conclusion and implication. By writing this dissertation, I realize that I only discovered the top of a huge iceberg and that my research has just started. Questions about optimizing the research methodology rise, new problems emerge, research ideas are born, and further research is needed to deal with recently discovered issues in the multilevel modeling of SSEDs. Literally a hundred directions for further research are possible, of which we will discuss the most important and timely ones.

10.1 Suggestion 1

As a first suggestion for further research, we suggest to conduct a systematic review of published and unpublished meta-analyses of SSEDs. This is needed to establish the empirical foundation for further studies by SSED researchers, meta-analysts and methodologists. In this dissertation, we chose a limited number of specific conditions to investigate. We could not focus on all interesting conditions, because of the limited extent of this dissertation. Also, the choice of the included conditions are based on a limited number of re-analyses of published and unpublised meta-analyses of SSED studies (e.g., Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011). A clear and more extended overview of published and unpublished meta-analyses together with an overview of meta-analystic and single-case data characteristics is missing in the literature. The overview can also report commonly encountered design complexities and combinations of complexities in the area of SSED research. We advise to formulate strict inclusion criteria about which meta-analyses of SSEDs to include, because a primary search for published meta-analyses of SSEDs using the social sciences citation index with the keywords ‘single-case’ or ‘single-subject’ or ‘interrupted time-series’ or ‘multiple-baseline’ in combination with ‘meta-analysis’ or ‘synthesis’ already resulted in 367 results. By refining the inclusion criteria and specifying “Education and Educational Research” as research area, and by only including studies published over the past 20 years, 65 results remain, which is more feasible to focus on. In addition to simply giving an extended overview of SSED data and synthesis characteristics, we suggest to retrieve the raw data graphically presented in the primary studies included in the meta-analyses. The raw data can be digitized using the statistical software program UnGraph (Biosoft, 2004). This results in a large dataset with which single-case analysts/synthesists can conduct a secondary analysis. By doing this, important information is

obtained and realistic values for a number of parameters can be defined. This is of importance for research synthesists and methodologists interested in improving the analysis and meta-analysis of SSEDs. The between-study variance, between-subject variance, within-subject variance, degree of autocorrelation, and covariance at the second and third level amongst other parameters can be estimated. Also trajectories during the baseline and treatment phase, design types, types of outcomes, number of data points per phase, phases per subject, number of subjects within the study, etc. can be determined. A sense of commonly encountered combinations of synthesis and SSED data characteristics can be obtained. Another interesting topic is the analysis technique used in the primary studies. We advise to keep track of the used analysis technique in order to identify the most commonly used analysis method (e.g., randomization tests, visual analysis, effect size estimates, etc.), and to investigate whether this is dependent on the year in which the article is published. This suggestion is timely, as in the past, the absence of this systematic approach resulted in not well documented conditions and complexities studied at the meta-analytic level of SSED studies.

10.2 Suggestion 2

There is a need to search for alternative estimation procedures to the maximum likelihood estimation in contexts of multilevel modeling of SSEDs that have the potential to solve the problem of biased variance estimates. In addition, estimation problems occur even for the fixed effects estimates when the multilevel model is getting more complex (i.e., when complex SSED data characteristics and synthesis characteristics are included). We suggest to investigate two alternative promising estimation procedures to synthesize SSED studies, namely Bayesian estimation paired with different prior distributions (Shadish et al., 2013) and parametric and nonparametric bootstrapping methods (Efron & Tibshirani, 1994; Mooney & Duval, 1993; Van den Noortgate & Onghena, 2005). This suggestion is timely and of crucial importance to solve estimation problems encountered when models get complex or when the research interest lies in (co-) variance estimation. Covariance estimation hardly got any attention in previous research, which is unfortunate because of the complementary and other source of information they can provide in addition to fixed effect estimates. For instance, the variance in estimated treatment effects (i.e., immediate treatment effect and treatment effect on time trend) between subjects and/or between studies can be estimated. There is a need to validate these alternative procedures using simulation studies, to program these procedures in

software packages, and to give clear guidelines in which circumstances which estimation procedure is advisable.

10.3 Suggestion 3

In this research, we focused on the validation of the basic three-level model and three extensions to it, namely standardizing the raw data, modeling external event effects and focusing on misspecification issues. However, we did not focus on extensions simultaneously modeled, for instance standardizing and modeling external event effects. Also, we only validated a limited number of extensions to the multilevel model, but suggested several extensions of which some are validated by simulation studies by other members of the research team. For instance, the modeling of non-linear trajectories is studied by Beretvas et al. (2013), the modeling of autocorrelation is investigated by Baek and Ferron (2013), and count data as outcome scores are modeled by Beretvas and Chu (2013). Ugille et al. (2012, 2013) focused on combining effect sizes across subjects and across studies, and validated a bias correction factor for the fixed effect estimates. If the SSED date are not graphically displayed in the primary studies, raw data cannot be retrieved, and then the approach of Ugille et al. (2012) is recommended. In addition, treatment effects estimates are often evaluated by group-comparison studies, in which data are aggregated over participants per condition, before comparing conditions. Results from these studies only permit assessment and explanation of variability between studies. Combining group-comparison and SSED data would result in more general treatment effect estimates, which is currently investigated by Ugille et al. (2014). The basic three-level model and extensions to it discussed in this dissertation only involved continuous outcome scales. However, the article of Shadish and Sullivan (2011), in which single-case characteristics of all 809 published single-case studies in 2008 were discussed, shows that nearly all outcome variables were some form of a count. Therefore, further research is needed to discuss the basic tree-level model and several extensions to it when the outcome scores are counts. For instance, a Poisson-distribution can be used. In addition, we discovered that if several studies are combined in the multilevel model, it might be the case that outcome scores are on different scales (e.g., interval, count, percentages). This influence of assuming continuous outcomes when in fact another type of scale is used, is an interesting research question. Also, estimating treatment and variance components across studies using different type of outcome scales is needed (or transferring outcome scales to continuous outcome scales could also be investigated). The systematic

review of published and unpublished meta-analyses of SSED studies (proposed as the first suggestion for further research) might provide some insight into which SSED data characteristics and synthesis characteristics are commonly encountered in the multilevel modeling of SSEDs. We advise to conduct a large simulation study, integrating realistic and complex conditions and parameter values for these conditions. These could be obtained by the secondary analysis of the retrieved meta-analytic SSED data (as presented in the first suggestion for further research). As these multilevel models are potentially getting complex, we advise to use alternative estimation procedures to the restricted maximum likelihood estimation, such as the Bayesian estimation procedure, which potentially results in less biased treatment and variance components estimates.

10.4 Suggestion 4

We recommend focusing on power in further research, as little is known about the power of the extended multilevel models to identify the statistical significance of treatment effect estimates and variance component estimates. The power is traditionally defined as the probability of correctly rejecting a false null hypothesis and is dependent on the sample size, the effect size, the probability of Type I error, and the specific experimental design (Howell, 2005). Although the first two simulation studies in this dissertation dealt with the issue of power, this was not the focus of interest in the other studies, neither was it the focus of the simulations conducted by the other members of our research team. Therefore, further research is needed to investigate to what degree the power is dependent on synthesis and SSED data characteristics. By examining the power, conditions can be determined under which the three-level model can be recommended. Furthermore, we suggest investigating under what conditions a reasonable power is obtained, and how the three-level modes can be modified and optimized to obtain a predefined power.

10.5 Suggestion 5

Standardization is of crucial importance when a researcher is interested in summarizing data across different studies, because it is likely that outcome scores from different studies are measured on different scales. In this dissertation, we proposed to standardize the SSED data for each subject separately before combining them across cases and across studies using the multilevel model. We validated the standardizing method proposed by Van den Noortgate and Onghena (2008), which involves dividing the subject-specific outcome scores by the

estimated within-case standard deviation prior to synthesis. However, the simulation study combining the standardized SSED data showed that the variance estimates are even more problematic in comparison to the three-level modeling of unstandardized raw data. We also discovered that the fixed effect estimates are underestimated when a large number of studies characterized with a small number of measurement occasions are included in the multilevel model. To standardize, we used the variance at the lowest level (the within-case variance), which is expected to be one. However, we estimated the (co)variances at the three levels and therefore, the estimated within-subject variance might slightly deviate from one. In the simulation studies, we found that the most problematic conditions are those where a lot of studies including a limited number of measurement occasions are included. When less measurement occasions are included, a less accurate within-subject variance might be obtained. Therefore, further research is needed to find a more optimal way of standardizing the SSED data. We suggest to evaluate whether analyzing standardized raw SSED data with a constrained level-1 variance would lead to less biased variance estimates. Another option is to use a fully Bayesian approach, and the use of a bias correction factor that was proposed by Hedges in another context (Ugille et al., 2013).

10.6 Suggestion 6

When using the multilevel model, specific assumptions about the data under investigation are made. In this dissertation, we only discussed one assumption in greater detail and that is the covariance specification at the second and the third level. We were especially interested in the influence of ignoring existing covariance on the fixed effects and variance component estimates. We did not focus on the level-1 residual specification as the research team of Prof. dr. John Ferron is investigating this. However, a combination of level-1, level-2 and level-3 variance misspecification issues is an interesting research line. Another assumption underlying the multilevel model is the assumption of normality. We assumed that the level-1 residuals come from a normal distribution and level-2 and level-3 residuals are multivariate normally distributed. However, single-case studies by definition focus on a small number of participants, which can negatively impact the distribution of the residuals. It would be interested to investigate to what degree the research results are accurate when we analyze non-normally distributed data. This can be accomplished by transforming the second and third level errors to a non-normal distribution, more specifically a distribution with heavier tails such as the *t*-distribution with small degrees of freedom, and a skewed distribution such as a

χ^2 distribution. Another assumption we made about the errors is that they are identically distributed. However, this might not be the case. For instance, in some SSED studies, it is obvious that the variance within the treatment phase is larger than the variance within the baseline phase. Also at the second level, it might be that the variance within one study is larger than the variance within another study. We did not take this into account. In all previous simulation studies investigating the multilevel modeling (Ferron et al., 2010; Moeyaert et al., 2013a, 2013b, 2013c, 2014b), designs were assumed to be balanced (homogeneous number of measurements within subjects, and subjects within studies, and homogeneous within- and between-subject and between-study variance). However, multilevel modeling data might be unbalanced. We suggest taking into account that the variance between subjects varies across studies, that the variance within subjects varies across subjects, and that the variance within studies varies between studies. We wonder whether this has a consequence on the fixed effects and variance components estimates.

10.7 Suggestion 7

Further research is needed to investigate cross-classified data in contexts of three-level modeling of SSED data. For instance measurements can be classified in a subject, but also in a setting. We already try to give some possible ways to deal with cross-classified data. One solution is to use a cross-classification mixed model. Another model that can be used is the multivariate three-level model, in which the dependent variables are the outcome scores in a particular setting; for instance Y_{hijk} is the outcome score in setting *h* for measurement occasion *i* from subject *j* and study *k*. A third option is including the setting as a measurement characteristic, by including the setting as a categorical predictor in the first level equation.

10.8 Suggestion 8

In this dissertation, we discussed that not only the number of published SSED studies is increasing at an astonishing rate, but also the number of meta-analyses (see Figure 1.7). Therefore, we argue that it might be interesting to include a predictor indicating the meta-analysis at the third level, but also this needs some further exploration as little is known about the modeling of predictors at the different levels of the multilevel model. For instance, the level two and level three equations can be extended by including predictors allowing us to investigate possible moderating effects of subject (e.g., age, gender, etc.) and study characteristics (e.g., study quality, publication year, etc.). We want to gain insight in how

many measurements, subjects and studies are required for adding a specific number of predictors at level two and level three to achieve a reasonable power.

10.9 Suggestion 9

Further work on multilevel modeling is needed to develop a practical manual about the multilevel modeling of SSED data in which all modeling options are discussed in detail, in accordance with the interpretation of the results obtained by the multilevel model. In this dissertation, we only discussed the possibilities of the statistical program SAS to combine single-case data across cases and across studies using the mixed procedure. However, the multilevel analysis is implemented in a variety of other statistical programs such as R, HLM, SPSS, and MLwiN. A translation of the SAS codes into other statistical software programs is needed, in accordance with their own advantages and disadvantages. In addition, developing a user-friendly tool in which the SSED researcher can explore their data might be helpful, as a number of different modeling options is possible with the multilevel model, and it might sometimes not be obvious which option to choose and to what degree the fixed effects and random effect estimates are dependent on these modeling options. In this way a quick sensitivity analysis could be conducted (i.e., to what extent are the results dependent on the modeling options). The underlying purpose of this dissertation is to guide single-case data analysts interested in using the multilevel model to summarize their data.

10.10 Suggestion 10

Previous research has focused on the coding schemes and syntheses of results from several types of SSEDs separately, including the multiple-baseline design, ABAB reversal design, ATD, and the changing criterion design (Moeyaert et al., 2014b; Shadish et al., 2013). However, a large proportion of actual SSEDs do not use a ‘pure’ design (such as a phase design or ATD), but rather a design that combines characteristics of two or more designs. Shadish and Sullivan (2011) found that 26% of the 809 studies they reviewed entailed combinations of the basic designs. As far as we know, no research has investigated coding, and effect size estimation for combinations of these designs.

10.11 Suggestion 11

In this dissertation we did not stress the importance to take specific issues related to SSED studies into account. A first issue is the need of stable baseline outcome scores prior to intervention in order to extrapolate accurately (Kazdin, 2011; Kratochwill et al., 2010). As a consequence, the start of the treatment phase cannot be determined a priori. Usually, single-case researchers apply a form of response-guided experimentation, where they continue to gather baseline data until an acceptable pattern emerges. Little is known about the consequences of response guided-experimentation on the treatment effect estimate across subjects and across studies. In this dissertation we were mainly interested in the immediate treatment effect and in the treatment effect on the slope. However, it might be the case that the effect of the treatment is delayed, and as a consequence, there is no linear increase in outcome scores during the treatment phase. Therefore, a non-significant immediate treatment effect and a significant change in slope can be found. It might be of more interest to identify the moment of intervention at which the treatment starts to have an impact on the outcome scores. Little is known about how to take delayed treatment effects into account and the consequences of ignoring this in the analysis. Another important issue that we did not elaborate on in this dissertation is the importance of randomization in SSEDs in order to increase the internal validity. Whenever it is possible, researchers should be advised to apply a form of randomization to eliminate alternative explanations for the change in outcome scores. Randomization can for instance be introduced for the start of the treatment, the assignment of measurement occasions to treatment phases, the number and order of phase repetitions, or the assignment of participants to baseline lengths. In further research, we recommend to implement a form of randomization in the design of the study and compare the results obtained by randomization tests and multilevel modeling or other methods. In contrast to the multilevel modeling approach suggested in this dissertation, randomization tests make no assumptions about the distribution of the residuals. An alternative approach to randomization is to include nonparametric bootstrap procedures in the multilevel modeling approach, which also needs further exploration.

10.12 Suggestion 12

In this dissertation, the focus of interest is the statistical analysis in order to summarize a large amount of data across subjects and across studies. However, visual analysis is also of importance as it can provide experimental control during the single-subject experiment. Both

procedures are complementary and necessary. Visual analysis techniques have long been acknowledged as effective and valuable (Michael, 1974). During visual analysis of the data, the effect of the independent variable and extraneous variables are evaluated while the SSED is being conducted. This ongoing process of data evaluation allows the applied SSED researcher to be responsive to the needs of the subject under investigation (Barlow & Hersen, 1984; Kazdin, 2011). For instance, the intervention can be adapted during observation or the intervention can be introduced only after a stable baseline pattern emerged. Further research is needed to provide guidelines about how both procedures can be implemented and how they can lead to consistent results.

10.13 Suggestion 13

Further research is needed to examine to what extent the results found here are informative to other research fields than education, such as in biomedical research fields. There is a need to look beyond our own applied research domain (i.e., educational research) and to learn from other research fields in which SSEDs are prevalent. Therefore, there is a need for a systematic review, giving an overview of published SSED studies per research field, and investigating to what extent the complexities and specific SSED data characteristics and synthesis characteristics are also found in other research fields. There is a need to work more interdisciplinary because the multilevel model is a generic approach and can be applied to synthesize SSED data in a variety of different research fields. Also, to my opinion, we could have learned a lot from the modeling of longitudinal data (Verbeke & Molenberghs, 2009), because they also deal with repeated measurements, and therefore cannot ignore the issue of autocorrelation. They also have to define trajectories, model different type of outcome scales, etc.

The Need for Further Research

We have plenty of other suggestions for further research, but we chose to only highlight the most important ones. This proves that there is still a lot of further research needed in the domain of single-subject designs and the multilevel modeling framework. We want to make clear that the research started in this dissertation is not finished and thus the dissertation should not be read as a closed book. We hope to inspire other researchers and motivate them to continue doing research in the field of single-subject and multilevel modeling.

REFERENCES |

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov , & F. Csaki (Eds), *Second International Symposium on Information Theory* (pp. 267-281). Budapest, Hungary: Akademiai Kiado.

Alen, E., Grietens, H., & Van den Noorgate, W. (2009). *Meta-analysis of single-case studies: an illustration for the treatment of anxiety disorders*. Unpublished master's thesis, Katholieke Universiteit Leuven, Leuven (Belgium).

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behavior Research Therapy, 31,* 621-631.

Anumendem, N. D., De Fraine, B., Onghena, P., & Van Damme, J. (2011). The impact of coding time on the estimation of school effects. *Quality and Quantity, 47,* 1021-1040. Doi: 10.1007/s11135-011-9581-3

Baek, E., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45,* 65-74.

Baek, E., Moeyaert, M., Petit-Bois, M., Beretvas, S.N., Van den Noortgate, W., & Ferron, J. M. (2013). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation*. Advance online publication. http://dx.doi.org/10.1080/09602011.2013.835740

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91-97.

Barlow, D. H., & Hayes, S.C. (1979). Alternating treatments design: one strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis, 12,* 199-210.

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Allyn & Bacon.

Beretvas, S. N., & Chu, Y. (2013). *Handling count data outcomes trajectories in multiple-baseline design studies*. Paper submitted for presentation at the annual meeting of the American Educational Research Association, Philadelphia, Pennsylvania.

Beretvas, S. N., & Chung, H. (2008). A review of single-subject experimental design meta-analyses: Methodological issues and practice. *Evidence-Based Communication and Assessment and Intervention, 2*, 129-141.

Beretvas, S. N., Hembry, I., Van den Noortgate, W., & Ferron, J. M. (2013). *Estimation of a nonlinear intervention phase trajectory for multiple baseline design data*. Manuscript submitted for publication.

Berkhof, J., & Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics, 29,* 201-218.

Biosoft (2004). UnGraph for Windows (Version 5.0). Cambridge, U.K.: Author.

Briesch, A. M., & Chafouleas, S. M. (2009). Review and analysis of literature on self-management interventions to promote appropriate classroom behaviors (1988-2008). *School Psychology Quarterly, 24*, 106-118. doi:10.1037/a0016159

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41*, 477-485. doi:10.3758/BRM.41.2.477

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill, & J. R. Levin (Eds.), *Single-case research design and analysis. New directions for Psychology and Education* (pp. 187–212). Mahwah, NJ: Erlbaum.

Busse, R. T., Kratochwill, T. R., & Elliott, S.N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269-285.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*, 387-400.

Chorpita, B. F., Albano, A., Heimberg, R.G., & Barlow, D.H. (1996). A symstematic replication of the prescriptive treatment of school refusal behavior in a single subject. *Journal of Behavior Therapy and Experimental Psychiatry, 27*, 281-290.

Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple-baseline designs. *Psychology in the Schools, 44*, 451-459.

Chung, Y. C., & Cannella-Malone, H. I. (2010). The effects of presession manipulations on automatically maintained challenging behavior and task responding. *Behavior Modification, 34*, 479-502. doi:10.1177/0145445510378380

Cohen, J. (Ed.). (1988). *Statistical power analysis for the behavioural sciences.* United states of America: Lawrence Erlbauw Associates.

Cools, W., Van den Noortgate, W., & Onghena P. (2008). ML-DEs: A program for designing efficient multilevel studies. *Behavior Research Methods*, *40*, 236-249.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: a step-by-step approach*. London: Sage.

Davis, D. H., Gagné, P., Frederick, L. D., Alberto, P. A., Rebecca, E. W., & Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behavior Modification, 37,* 62-89.

Denis, J., Van den Noortgate, W., & Maes, B. (2011). Self-injurious behavior in people with profound intellectual disabilities: A meta-analysis of single-case studies. *Research in Developmental Disabilities, 32*, 911-923.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573-579.

Edgington, E. S. (1967). Statistical inference from N = 1 experiments. *Journal of Psychology, 65*, 195-199.

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational and Behavioral Statistics, 5*, 235-251.

Edgington, E. S., & Onghena, P. (2007). *Randomization Tests* (4th ed.). London: Chapman & Hall.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Washington, D.C.: Chapman & Hall/crc.

Fai, A. H.-T., & Cornelius, P. L. (1996). Approximate *F*-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation & Simulation*, *54*, 363-378.

Farmer, J., Owens, C. M., Ferron, J.M., & Allsopp, D. (2010*). A review of social science single-case meta-analyses.* Manuscript in preparation.

Ferron, J. M., Bell, B. A., Hess, M. F., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: the utility of multilevel modeling approaches. *Behavior Research Methods, 41*, 372-384.

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilvel-modeling approaches. *Behavior Research Methods, 42*, 930-943.

Ferron, J. M., & Jones, P.K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education, 75*, 66-81.

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014*).* Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*. Manuscript accepted for publication.

Ferron, J. M., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education, 64*, 231-239.

Ferron, J. M., & Scott, H. (2005). Multiple baseline designs. In B. Everitt & D. Howell (Eds). *Encyclopedia of Behavioral Statistics* (Vol. 3, pp. 1306-1309). West Sussex, UK: Wiley & Sons Ltd.

Ferron, J.M., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education, 70*, 165-178.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes, 54*, 137-154.

Fouladi, R. T., & Shieh, Y. (2004). A comparison of two general approaches to mixed model longitudinal analyses under small sample size conditions. *Communications in Statistics: Simulation and Computation*, *33*, 807–824.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1997). Introduction. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 1–12). Mahwah, NJ: Erlbaum.

Gast, D. L. (2010). Applied research in education and behavioral sciences. In D. L. Gast (Ed.), *Single-subject research methodology in behavioral sciences* (pp. 1–19). New York, NY: Routledge.

Gentile, J., Roden, A., & Klein, R. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 5*, 193-198.

Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *Journal of Applied Behavioral Science, 25*, 298-314.

Giesbrecht, F. G., & Burns, J. C. (1985). Two-stage analysis based on a mixed model: large sample asymptotic theory and small-sample simulation results. *Biometrics*, *41*, 477-486.

Gingerich, W. J. (1984). Meta-Analysis of applied time-series data. *The Journal of Applied Behavioral Science, 20*, 71-79.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Goldstein, H. (1995). *Multilevel statistical models*. London, England: Edward Arnold.

Goldstein, H., Healey, M. J. R., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine, 13*, 1643-1655.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.

Guralnick, M. J. (1978). The application of single-subject research designs to the field of learning disabilities. *Journal of Learning Disabilities, 11*, 415-421.

Gomez, E., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communication in Statistics: Simulation and Computation, 34*, 377–392.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with the application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355-370.

Hanley, G. P., Iwata, B. A., Thompson, R. H., & Lindberg, J. S. (2000). A component analysis of "stereotypy as reinforcement" for alternative behavior. *Journal of Applied Behavior Analysis, 33*, 285-297.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*, 224-239.

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65. doi:10.1002/jrsm.5

Heyvaert, M., Maes, B., Van Den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33*, 766-780.

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2014). Systematic review of restraint interventions for challenging behaviour among persons with intellectual disabilities: Focus on effectiveness in single-case experiments. *Journal of Applied Research in Intellectual Disabilities*, Advanced Online Publication. doi:10.1111/jar.12094

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. *Sociological Methods & Research, 26,* 329-367.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165-179.

Horner, R., H. & Spaulding, S. (in press). Single-Case Research Designs. Encyclopedia. Springer.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children, 35,* 269-290.

Howell, D.C. (2005). Power. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1558–1564). Chichester: Wiley.

Hox, J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Erlbaum.

Huitema, B. E., & McKean, J. W. (1994). Two biased-reduced autocorrelation estimators: rF1 and rF2. *Perceptual and Motor Skills, 78,* 323–330.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 1,* 104-116.

Huitema, B. E. & McKean, J.W. (2000). Design specification issues in time series intervention models. *Educational and Psychological Measurement, 60,* 38-58.

Ittenbach, R. F., & Lawhead, W. F. (1997). Historical and philosophical foundations of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 13–39). Mahwah, NJ: Erlbaum.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42,* 805 - 820.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44,* 483-493.

Jones, R. J., Weinrott, M. R., & Vaught, R. S., (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277-283.

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association, 79,* 853-862.

Kahng, S. W., Chung, K. M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 43*, 35-45.

Kalaian, H. A., & Raudenbush, S.W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1*, 227-235.

Kazdin, A.E. (1982) *Single case research designs, methods for clinical and applied settings.* Oxford: University Press.

Kazdin, A.E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Kazdin, A. E., & Kopel, S. A. (1975). On resolving ambiguities of the multiple-baseline design: problems and recommendations. *Behavior Therapy, 6*(5), 601-608.

Kennedy, M. M. (1979). Generalizing from single case studies. *Evaluation Quarterly, 3*, 661-678.

Kennedy, C. H. (2005). *Single-case designs for educational research*. New York: Allyn and Bacon.

Kinugasa, T., Cerin, E., & Hooper, S. (2004). Single-subject research designs and data analyses for assessing elite athletes' conditioning. *Sports Medicine, 34*, 1035-1050.

Koegel, L. K. Camarata, S. M. Valdez-Menchaca, M. ,& Koegel, R. L. (1998). Setting generalization of question-asking by children with autism. *American Journal on Mental Retardation*, *102*, 346-357.

Koegel, R. L., Symon, J. B., & Koegel, L. K. (2002). Parent education for families of children with autism living in geographically distant areas. *Journal of Positive Behavior Interventions, 4*, 88-103.

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: a potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods, 3*, 206-217.

Koehler, M. J., & Levin, J. R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments & Computers, 32*, 367-371.

Kokina, A., & Kern, L. (2010). Social story interventions for students with autism spectrum disorders: a meta-analysis. *Journal of Autism and Developmental Disorders, 40*, 812-826.

Koutsoftas, A. D., Harmon, M. T., & Gray, S. (2009). The effect of tier 2 intervention for phonemic awareness in a response-to-intervention model in low-income preschool classrooms. *Language, Speech, and Hearing Services in Schools, 40*, 116–130.

Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement, 64,* 224–242.

Kratochwill, T. R. (1985). Case study research in school psychology. *School Psychology Review, 14,* 204–215.

Kratochwill, T. R., Alden, K., Demuth, D., Dawson, D., Paicucci, C., Arntson, P., et al. (1974). A futher consideration in the application of an analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 7,* 629-633.

Kratochwill, T. R., & Brody, G.H. (1978). Single subject designs. A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2,* 291-307.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J.R. (1992). *Single-case research design and analysis: New directions for psychology and education.* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15,* 124–144. doi:10.1037/a0017736

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65,* 73-93.

Kwok, O., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research, 42,* 557–592.

Lambert, M.C, Cartledge, G., Heward, W.L., & Lo, Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions, 8,* 86-99.

Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis, 4,* 391-400.

LeBlanc, L. A., Geiger, K. B., Sautter, R. A., & Sidener, T. M. (2007). Using the natural language paradigm (NLP) to increase vocalizations of older adults with cognitive impairments. *Research in Developmental Disabilities*, *28*, 437-444.

Lenz, A. S. (2013). Calculating effect size in single-case research: A comparison of nonoverlap methods. *Measurement and Evaluation in Counseling and Development, 46*, 64-73.

Levin, J. R., O'Donnell, A. M., & Kratochwill, T. R. (2003). Educational/psychological intervention research. In I. B. Weiner (Series Ed.) and W. M. Reynolds & G. E. Miller (Vol. Eds.). *Handbook of psychology: Vol. 7. Educational psychology* (pp. 557–581). Hoboken, NY: Wiley.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS© system for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Lindberg, J. S., Iwata, B. A., & Kahng, S. W. (1999). On the relation between object manipulation and stereotypic self-injurious behavior. *Journal of Applied Behavior Analysis*, *32*, 51-62.

Luiselli, J.K., Suskin, L., & McPhee, D.F. (1981). Continuous and intermittent application of overcorrection in a self-injurious autistic child: Alternating treatments design analysis, *Journal of Behavior Therapy and Experimental Psychiatry, 12*, 355-358.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49,* 301-321. doi:10.1016/j.jsp.2011.03/004

Manolov, R., & Solanas, A. (2009). Factors affecting visual inference in single-case designs. *Spanish Journal of Psychology, 12*, 823-832.

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology, 51*, 201-215.

Mastropieri, M., & Scruggs, T. (1985). Early intervention for socially withdrawn children. *The Journal of Special Education, 19*, 429-441. doi:10.1177/002246698501900407

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341-351.

McCord, B. E., Grosser, J. W., Iwata, B. A., & Powers, L. A. (2005). An analysis of response-blocking parameters in the prevention of pica. *Journal of Applied Behavior Analysis, 38,* 391-394. doi:10.1901/jaba.2005.92-04

McGoey, K. E., & DuPaul, G. J. (2000). Token reinforcement and response cost procedures: Reducing the disruptive behavior of preschool children with Attention Deficit/Hyperactivity Disorder. *School Psychology Quarterly, 15,* 330-343. doi:10.1037/h0088790

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 5,* 87-101.

McLean, R. A., Sanders, W. L., & Stroup,W. W. (1991), A unified approach to mixed linear models. *The American Statistician*, *45,* 54–64.

McReynolds, L.V., & Thomspon, C.K. (1986). Flexibility of single-subject experimental designs. Part I. *Journal of Speech and Hearing Disorders*, *51*, 194-203.

Methe, S.A., Kilgus, S.P., Nieman, C., & Riley-Tillman, T.C. (2012). Meta-analysis of interventions for basic mathematics computation in single-case research. *Journal of Behavioral Education, 21*, 230-253.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse, *Journal of Applied Behavior Analysis*, *7*, 647-653.

Michielutte, R., Shelton, B., Paskett, E.D., Tatum, C.M., & Velez, R. (2000). Use of an interrupted time-series design to evaluate a cancer screening program. *Health Education Research*, *15*, 615-623.

Moes, D. R. (1998). Integrating choice-making opportunities within teacher-assigned academic tasks to facilitate the performance of children with autism. *The Association for Persons with Severe Handicap, 23*, 319-328.

Moeyaert, M., Bunuan, R., & Beretvas, S. N. (2014). *Multilevel meta-analysis of alternating treatment design studies: A Monte Carlo simulation study*. Paper submitted for presentation at the annual meeting of the American Educational Research Association, Philadelphia, Pennsylvania.

Moeyaert, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology, 52,* 191-211. doi: http://dx.doi.org/10.1016/j.jsp.2013.11.003

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2013a). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education, 82,* 1-21. doi: 10.1080/00220973.2012.745470

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2013b). Three-level analysis of standardized single-case experimental data: Empirical validation. *Multivariate Behavior Research, 48,* 719-748. doi: 10.1080/00273171.2013.816621

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, T., & Van den Noortgate, W. (2013c). Modeling external events in the three-level analysis of multiple-baseline across participants designs: A simulation study. *Behavior Research Methods, 45,* 547-559. doi: 10.3758/s13428-012-0274-1

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014a). *The misspecification of the covariance structures in multilevel models for single-case data: A Monte Carlo simulation study.* Manuscript submitted for publication.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014b). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental design research. *Behavior Modification.* Manuscript accepted for publication.

Moeyaert, M., Ugille, M., Ferron, Onghena, P. J., Heyvaert, M., Beretvas, S. N., & Van den Noortgate, W. (2014c). Estimating intervention effects across different types of single-subject experimental designs: Empirical Illustration. *School Psychology Quarterly*, 52 (2).

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical Inference.* Sage Publications, Beverly Hills, CA

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist, 56,* 119-127.

Murphy, D. L., & Pituch, K. A. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education, 77,* 255-282.

Nagler, E., Rindskopf, D., & Shadish, W. (2008). *Analyzing data from small N designs using multilevel models: A procedural handbook*. Unpublished manuscript.

National Research Council (2002). Committee on Scientific Principles for Education Research. Center for Education. Division of Behavioral and Social Sciences and Education. In R. J. Shavelson & L. Towne (Eds.), *Scientific research in education.* Washington, DC: National Academy Press.

Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification, 30*, 295-314.

Nugent, W. R. (1996). Integrating single-case and group-comparison designs for evaluation research. *Journal of Applied Behavioral Science, 32*, 209-226.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71*, 137-148.

Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment, 14*, 153-171.

Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1850-1854). Chichester: Wiley.

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research & Therapy, 32*, 783-786.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21*, 56-68.

Ottenbacher, K. J. (1992). Analysis of data in idiographic research. *American Journal of Physical Medicine & Rehabilitation, 71*, 202-208.

Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods, 44,* 795-805.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*, 194-204.

Parker, R. I., & Vannest, K. (2008). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*, 357-367.

Parker, R. A., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303-322

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22*, 109-116.

Petit-Bois, M., Beak, E. K., & Ferron, J. M. *Consequences of misspecification of growth trajectories when meta-analyzing single-case data using a three-level model.* Paper presented at the American Educational Research Association conference, Vancouver, British Columbia, Canada, 13-17 April 2012.

Petit-Bois, M., Baek, E. K., & Ferron, J. M. (2013, April). *The effect of error structure specification on the meta-analysis of single-case studies: A Monte Carlo study*. Poster presented at the American Educational Research Association Annual Meeting, San Francisco, CA.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods. Second edition* (Vol. 1). London, New Delhi.

Rindskopf, D. (2013). A Bayesian estimate of d in single case designs. Paper submitted for presentation at the annual meeting of the American Educational Research Association, San Francisco, California.

Rindskopf, D., & Ferron, J. (in press). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.) *Single-case intervention research: Methodological and data-analysis advances* (pp. XXX-XXX). American Psychological Association.

Roane, H. S., Piazza, C. C., Sgro, G. M., Volkert, V. M., & Anderson, C. M. (2001). Analysis of aberrant behaviour associated with Rett syndrome. *Disability and Rehabilitation, 23*(3-4), 139-148.

Rolider, A., Williams, L., Cummings, A., & Van Houten, R. (1991). The use of a brief movement restriction procedure to eliminate severe inappropriate behavior. *Journal of Behavior Therapy and Experimental Psychiatry*, 22, 23-30. doi:10.1016/0005-7916(91)90029-5

Roscoe, E. M., Iwata, B. A., & Goh, H.-L. (1998). A comparison of noncontingent reinforcement and sensory extinction as treatments for self-injurious behavior. *Journal of Applied Behavior Analysis*, *31*, 635-646.

Salzberg, C. L., Strain, P. S., & Baer, D. M. (1987). Meta-analysis for single subject research: When does it clarify, when does it obscure. *Remedial and Special Education*, *8*, 140-146.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6,* 309-316. doi:10.1007/BF02288586

Schreibman, L., Stahmer, A. C., Barlett, V. C., & Dufek, S. (2009). Brief report: toward refinement of a predictive behavioral profile for treatment outcome in children with autism. *Research in Autism Spectrum Disorders, 3,* 163-172

Schwarz, G. (1978). Estimating the dimension of a model. *Annuals of Statistics, 6,* 461-464.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research methodology: Methodology and validation. *Remedial and Special Education, 8,* 24-33.

Shadish, W.R., Brasil, I.C.C., Illingworht, D.A., White, K.D., Galindo, R., Nagler, E.D., & Rindskopf, D.M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods, 41,* 177-183.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: new applications and some agenda items for future research. *Psychological Methods, 18,* 385-405.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113,* 95-109.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2,* 188-196.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V., & Sullivan, K.J. (2012). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods, 45,* 813-821.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43,* 971-980.

Sherer, M. R., & Schreibman, L. (2005). Individual behavioral profiles and predictors of treatment effectiveness for children with autism. *Journal of Consulting and Clinical Psychology, 73,* 525-538.

Shogren, K. A., Fagella-Luby, M. N., Bae, J. S., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior. *Journal of Positive Behavior Interventions, 6,* 228-237.

Singer, J. D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Model change and event occurrence.* Oxford: Oxford University Press.

Snijders, T., & Bosker, R. J. (1993). Standard errors and sample Sizes for two-level research. *Journal of Educational and Behavioral Statistics, 18,* 237-259.

Snijders, T., & Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage

Strube, M. J., Gardner, W., & Hartmann, D. P. (1985). Limitations, liabilities and obstacles in reviews of the literature: The current status of meta-analysis. *Journal of Consulting and Clinical Psychology, 5,* 63–68.

Swanson, H. L., & Sachse-Lee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities, 33,* 114-136.

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the Single-Case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation, 18,* 385–401.

Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education.* Columbus, OH: Merrill.

Thompson, R. H., Iwata, B. A., Conners, J., & Roscoe, E. M. (1999). Effects of reinforcement for alternative behavior during punishment of self-injury. *Journal of Applied Behavior Analysis, 32,* 317-328.

Thorp, D. M., Stahmer, A. C., & Schreibman, L. (1995). The effects of sociodramatic play training on children with autism. *Journal of Autism and Developmental Disorders, 25,* 265-282.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46,* 234-240.

Tummers, B. (2005-2006). DataThief III manual v. 1.1. Retrieved from http://www.datathief.org/DatathiefManual.pdf

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noorgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44,* 1244-1254.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van Den Noortgate, W. (2013). Bias corrections for standardized effect size estimates used with single-subject experimental designs. *Journal of Experimental Education.* Advance online publication. doi: 10.1080/00220973.2013.813366

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noorgate, W. (2014). *Combining group-comparison and single-subject experimental designs.* Manuscript sumbitted for publication.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325-346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*, 1-10.

Van Den Noortgate, W., Onghena, P. (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods, Instruments & Computers, 37*, 11-22.

Van den Noortgate, W., Opdenakker, M., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement, 16*, 281-303.

Van den Noortgate, W., Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*, 196-209.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence Based Communication Assessment and Intervention, 2*, 142-151.

Verbeke, G., Molenberghs, G. (2009). *Linear mixed models for longitudinal data (Paperback Edition).* New York: Springer.

Velicer, W. F., & Fava, J. L. (2003). *Time series analysis.* In J. Schinka & W. F. Velicer (Eds.), *Research methods in psychology* (pp. 581-606). New York: John Wiley.

Wacker, D. P., Steege, M., & Berg, W. K. (1988). Use of single-case designs to evaluate manipulable influences on school performance. *School Psychology Review, 17*, 651–657.

Wampold, B.E., & Worsham, N.L. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment, 8*, 135-143.

Wang, S., Cui, Y., & Parrila, R. (2011). Examining the effectiveness of peer-mediated and video-modeling social skills interventions for children with autism spectrum disorders: a meta-analysis in single-case research using HLM. *Research in Autism Spectrum Disorders, 5*, 562-569.

Wang, J., Xie, H., & Fisher, J.H. (2012). *Multilevel models: Applications using SA*S. Berlin, Germany: Higher Education Press and Walter de Gruyter GmbH & Co. KG.

White, O. R. (1987). The quantitative synthesis of single-subject research: Methodology and validation. Comment. *Remedial and Special Education, 8*, 34–39.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual-subject research. *Behavioral Assessment, 11*, 281-296.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single subject data. *Journal of Special Education, 44*, 18–28.

Wolfinger, R. D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*, 205-230.

Zhou, L. M., Goff, G. A., & Iwata, B. A. (2000). Effects of increased response effort on self-injury and object manipulation as competing responses. *Journal of Applied Behavior Analysis, 33*(1), 29-40.

ADDENDA |

Addendum A: SAS codes

Addendum A1: SAS code used to standardize single-case data in (Chapter 3)

We prepare a dataset called 'raw' with the raw data, and then we run following codes in SAS 9.3:

- **Step 1**: we conduct an ordinary least square regression analysis on the raw single-case data in order to estimate the residual standard deviation per subject. In this example, the data file 'raw' contains the raw unstandardized data. We give the dataset containing the parameter estimates the name 'uncorrected' and the dataset containing the RMSE (= root mean squared error) per case the name 'MSE'.

```
PROC REG DATA=raw;
        BY study case;
        MODEL y = t D Dt;
         ODS OUTPUT ParameterEstimates=uncorrected anova=MSE;
RUN;
```

- **Step 2**: we calculate the standard deviation of $\hat{e}_{ijk}$, which is the RMSE. The new dataset 'MSE' only contains the study and subject number and the subject specific RMSE.

```
DATA MSE;
        SET MSE;
        WHERE source='Error';
        RMSE=sqrt(MS);
        KEEP study case RMSE;
RUN;
```

- **Step 3**: we merge the estimated residual within-subject standard deviation (the RMSE in the dataset named 'MSE') with the raw estimated data ('uncorrected').

```
DATA uncorrected;
        MERGE uncorrected MSE;
        BY study case;
RUN;
```

- **Step 4**: we calculate the standardized scores by dividing the scores (Y) by the estimated residual within-subject standard deviation (RMSE)

```
DATA raw;
        MERGE raw MSE;
        BY study case;
        yS=y/RMSE;
RUN;
```

- **Step 5 (three-level analysis)**: we use PROC MIXED in order to estimate the treatment effects over cases and over studies using the standardized single-case data (yS from the dataset 'raw'). We apply the PROC MIXED procedure on the dataset 'raw'. In the second statement we use study and case as CLASS variables which indicates that these variables are categorical. In the third statement we use MODEL to indicate the fixed part. The variable Ys (the standardized single-case data) is the dependent variable and the variables t (time), D (condition), Dt (interaction between time and condition) are the independent variables. The model estimates by default the intercept. If you are interested in the estimation of the treatment effects over cases and over studies using the unstandardized data, then you just use Y instead of Ys in the MODEL statement. The next statement includes the random part of the model, using RANDOM. We define that the intercept, t, D and Dt can vary randomly vary across studies (SUB = study) and across cases [SUB = case(study)].

```
PROC MIXED DATA=raw;
        CLASS study case;
        MODEL yS=t D Dt/ SOLUTION;
        RANDOM intercept t D Dt/ SUB=study;
        RANDOM intercept t D Dt/ SUB=case(study);
RUN;
```

Addendum A2: SAS code multilevel analysis of unstandardized single-case data modeling and ignoring covariance at level 2 and level 3 (Chapter 5)

SAS code used for the empirical example.
We prepared the dataset called 'raw' and then we ran following programs in SAS 9.3:

1. **Three-Level Analysis Ignoring Covariance**

Code
We use PROC MIXED in order to estimate the treatment effects over cases and over studies.

```
PROC MIXED DATA = raw;
CLASS   study   case;
MODEL   Y   =   t   D   Dt   /   SOLUTION DDFM = sat;
RANDOM   intercept   t   D   Dt   /   SUB = study;
RANDOM   intercept   t   D   Dt   /   SUB = case(study);
ODS OUTPUT   solution = fixed1 covparms = random1;
RUN;
```

In the first statement, we call the PROC MIXED procedure. Then we use DATA= statement to indicate which dataset we will use. In the second statement we use study and case as CLASS variables. This means that we define study and case as categorical variables. In the third statement we use MODEL to indicate the fixed part. The variable Y is the dependent variable and the variables *t* (time), *D* (condition), and *Dt*(interaction between *t*, centered around its value at the start of the treatment phase, and *D)* are the independent variables. The model estimates by default the intercept. The statement also involves the command SOLUTION. This is to request the estimates, standard errors, t-values, and *p*-values for the average intercept, and effects of *t*, *D* and *Dt* (the fixed effects or independent variables). DDFM = sat indicates that the Satterthwaite method was used to estimate the degrees of freedom. The next statement includes the random part of the model, using RANDOM. We define that the intercept, and the effects of *t*, *D* and *Dt* can vary randomly across studies (SUB = study) and across cases, which are nested in studies [SUB = case(study)]. The fixed effects estimates and the random effects estimates are saved respectively in the data files fixed1 and random1.

Output

Fixed1
The average intercept, and the average effects of *t*, *D* and *Dt* across cases and across studies are estimated.

	Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
1	Intercept	3.0991	0.9122	18.3	3.40	0.0031
2	T	0.02156	0.04306	12.9	0.50	0.6249
3	D	-2.6628	0.6646	18.4	-4.01	0.0008
4	Dt	-0.04448	0.04406	12.5	-1.01	0.3318

Random1
In the first row, the estimated between-study variance of the intercept is presented. The second, third and fourth row represent the between-study variance of the trend during baseline, of the immediate treatment effect, and of the treatment effect on the time trend respectively. The next four rows present the between-case variance of the intercept, the trend during baseline, the immediate treatment effect and the treatment effect on the time trend. The remaining row contains the estimated within-case variance.

	Cov Parm	Subject	Estimate
1	Intercept	study	13.5166
2	T	study	0.01858
3	D	study	5.8875
4	Dt	study	0.01716
5	Intercept	case(study)	2.3916
6	T	case(study)	0.002461
7	D	case(study)	2.3831
8	Dt	case(study)	0.000294
9	Residual		1.0885

2. Three-Level Analysis Modeling Covariance

We use again the SAS PROC MIXED procedure. But now we have to specify that the covariance matrix has to be estimated. We can accomplish that by using the 'TYPE' statement: TYPE = un in the random part of the model. This means that we estimate the variances and covariance. In the simulation study, we only estimated the covariance between the regression coefficients of the treatment effects. This is reasonable, because we set the intercept and the trend during the baseline level on zero (to obtain a more clear interpretation of the treatment effects). Therefore we restrained some parameters to zero using the PARMS statement in line 5. Here you can see that parameter 2, 4, 5, 7, 8, 12, 14, 15, 17 and 18 of the covariance matrix are restrained to zero. This means that the covariance between the effect of *t* and the intercept UN(2,1), between the effect of *D* and the intercept UN(3,1), between the effects of *D* and *t* UN(3,2), between the interaction effect of *Dt* and the intercept UN(4,1), between the effects of *Dt* and *t* (4,2) are restrained to zero at the second level (SUB = study) and at the third level [SUB = case(study)].

Code

```
PROC MIXED DATA = raw;
CLASS    study    case;
MODEL Y =   t   D   Dt   /   SOLUTION;
RANDOM   intercept   t   D   Dt   /   SUB = study   TYPE = un;
RANDOM   intercept   t   D   Dt   /   SUB = case(study)   TYPE=un;
PARMS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1   /   HOLD=2 4 5 7 8 12 14 15 17 18;
ods output   solution = fixed2 covparms = random2;
RUN;
```

Output

Fixed2

	Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
1	Intercept	2.9541	0.8951	18.2	3.30	0.0039
2	T	0.07311	0.05484	17.5	1.33	0.1995
3	D	-2.9961	0.7087	18	-4.23	0.0005
4	Dt	-0.1004	0.05216	17	-1.92	0.0712

Random2

	Cov Parm	Subject	Estimate
1	UN(1,1)	study	12.9478
2	UN(2,1)	study	0
3	UN(2,2)	study	0
4	UN(3,1)	study	0
5	UN(3,2)	study	0
6	UN(3,3)	study	7.6500
7	UN(4,1)	study	0
8	UN(4,2)	study	0
9	UN(4,3)	study	-0.1299
10	UN(4,4)	study	0.01800
11	UN(1,1)	case(study)	2.3323
12	UN(2,1)	case(study)	0
13	UN(2,2)	case(study)	0.05844
14	UN(3,1)	case(study)	0
15	UN(3,2)	case(study)	0
16	UN(3,3)	case(study)	3.5064
17	UN(4,1)	case(study)	0
18	UN(4,2)	case(study)	0
19	UN(4,3)	case(study)	0.3382
20	UN(4,4)	case(study)	0.04411
21	Residual		1.0181

Addendum A3: SAS code single-level analysis (Chapter 6)

We give a description of the SAS code that can be used to conduct a separate linear regression for each participant of the multiple-baseline study of Laski et al. (1988) using design matrix 1. The same steps are used for design matrix 2, design matrix 3 and design matrix 4. The only difference between the design matrices is the coding of the time variable and multiple time variables are used in design matrix 2 and design matrix 4.

1. **SAS Code Design Matrix 1**

In the first statement, we conduct an ordinary least square regression analysis using the dataset named "Laski". In the second statement we use a "by" statement to indicate that we conduct the regression analysis for each case separately. In the third statement we define the model. The variable Y is the dependent variable and the variables *Time*, *treatment*, and *TreatmentTime1* are the independent variables. By default, the intercept is estimated.

```
PROC REG DATA=Laski;
        BY case;
        MODEL y = Time    Treatment     TreatmentTime1;
RUN;
```

2. **SAS Code Design Matrix 2**

```
PROC REG DATA= Laski;
        BY case;
        MODEL y = Time    Treatment    TreatmentTime;
RUN;
```

3. **SAS Code Design Matrix 3**

```
PROC REG DATA=Laski;
        BY case;
        MODEL y = Time1    Treatment    TreatmentTime1;
RUN;
```

4. **SAS Code Design Matrix 4**

```
PROC REG DATA=Laski;
        BY case;
        MODEL y = Time2    Treatment    TreatmentTime1;
RUN;
```

Addendum A4: SAS code two-level analysis (Chapter 6)

We provide a description of the SAS code that can be used to conduct a two-level analysis for the multiple-baseline design data from Laski et al. (1988).
In the first statement, we request a two-level analysis on the dataset named "Laski". In the second statement we specify case as a CLASS variable which indicates that this variable is categorical. In the third statement we use MODEL to indicate the fixed effects in the model. The variable Y is the dependent variable and the variables *Time*, *Treatment*, and *TreatmentTime1* are the independent variables. By default, an intercept is estimated. The next statement specifies the random effects in the model, using RANDOM. We specify, here, that the *intercept*, *Time*, *Treatment*, and *Treatment Time1* can vary randomly across cases (SUB = case). By using the statement SOLUTION in the random statement, the case-specific regression coefficients are estimated using empirical Bayes estimation.

```
PROC MIXED DATA=Laski;
        CLASS case;
        MODEL Y= Time    Treatment    Treatment Time1/ SOLUTION ;
        RANDOM intercept    Time    Treatment    Treatment Time1/ SOLUTION SUB=case;
RUN;
```

Addendum A5: SAS code two-level analysis (Chapter 7)

1. Model 1: Basic Two-Level Model

Model 1A

```
PROC MIXED  COVTEST DATA=TwoLevel METHOD=ML;
        CLASS case;
        MODEL Y =  Phase /  SOLUTION DDFM=sat;
        RANDOM Intercept  Phase /  SUB=Case;
        ODS OUTPUT solutionF=fixed1a covparms=random1a fitstatistics=fit1a;
RUN;
```

Model 1B

```
PROC MIXED  COVTEST DATA=TwoLevel METHOD=ML;
        CLASS Case;
        MODEL Y =  A1B1 B1A2 A2B2  /  SOLUTION DDFM=sat;
        RANDOM Intercept A1B1 B1A2 A2B2 /  SUB=Case;
        ODS OUTPUT solutionF=fixed1b covparms=random1b fitstatistics=fit1b;
RUN;
```

In the first statement, the mixed procedure is called. The DATA = statement refers to the data set in which the data are stored. The METHOD = statement asks the maximum likelihood estimation procedure. In the second line, the variable case, identifying the cases, is defined as a categorical variable. In the third line, the fixed part of the model is described. The variable Y is defined as the dependent variable and the variable Phase in Model A and the variables A1B1 B1A2 A2B2 in Model B are defined as independent variables. The model includes an intercept by default. The SOLUTION-option is used to request in the output the estimates, standard errors, *t*-statistics and *p*-values for significance testing for all fixed effects. The RANDOM statement is used to describe the random part of the model. We indicate that the intercept and phase can vary randomly across cases. If one is interested in the case-specific baseline levels and treatment effects, the code can be adapted by including the SOLUTION-option in the random part. The ODS OUTPUT is used to save the fixed effect estimates (solutionF), the random effect estimates (covparms), and the fit statistics (fitstatistics) in output files.

2. Model 2: Modeling Autocorrelation and Heterogeneous Within-Case Variance

Model 2a

```
PROC MIXED  covtest DATA=TwoLevel METHOD=ML;
        CLASS Case Phase;
        MODEL Y =  Phase /  SOLUTION DDFM=sat;
        RANDOM Intercept  Phase /  SUB=Case;
        REPEATED / SUB=case GROUP=Phase TYPE=ar(1);
        ODS OUTPUT solutionF=fixed2a covparms=random2a fitstatistics=fit2a;
RUN;
```

Model 2b

```
PROC MIXED  covtest DATA=TwoLevel METHOD=ML;
        CLASS Case Phase;
        MODEL Y =  A1B1 B1A2 A2B2  /  SOLUTION ddfm=sat;
        RANDOM Intercept  A1B1 B1A2 A2B2  /  SUB=Case;
        REPEATED / SUB=Case GROUP=Phase TYPE=ar(1);
        ODS OUTPUT solutionF=fixed2b covparms=random2b fitstatistics=fit2b;
RUN;
```

Compared with the former programs (Model 1 and Model 2) there is an additional line requesting the modeling of a first order autocorrelation within cases. This random part on the first level is modeled using the repeated statement. The option type = AR(1) requests modeling a first-order autocorrelation within cases.

3. Model 3: Autocorrelation + Heterogeneous Within-Case Variance + Linear Time Trend in the Treatment Phase

Model 3

```
PROC MIXED  COVTEST DATA=TwoLevel METHOD=ML;
        CLASS Case Phase;
        MODEL Y = A1B1 B1A2 A2B2  T1 A1B1*T2 B1A2*T3 A2B2*T4
        / SOLUTION DFM=sat;
        RANDOM Intercept A1B1 B1A2 A2B2 T1 A1B1*T2 B1A2*T3 A2B2*T4
        /SUB=Case;
        REPEATED / SUB=Case GROUP=Phase TYPE=ar(1);
        ODS OUTPUT solutionF=fixed3 covparms=random3 fitstatistics=fit3;
RUN;
```

Compared with Model 3, the model specification accordingly with the random part is changed by adding time variables.

4. Model 4: Autocorrelation + Heterogeneous Within-Case Variance + Linear Time Trend in the Treatment Phase + Level Two Predictor

Model 4

```
PROC MIXED  COVTEST DATA=TwoLevel METHOD=ML;
        CLASS Case Phase Class ;
        MODEL Y =  class A1B1 B1A2 A2B2  T1 A1B1*T2 B1A2*T3 A2B2*T4/
        SOLUTION  DDFM=sat;
        RANDOM Intercept  A1B1 B1A2 A2B2  T1 A1B1*T2 B1A2*T3 A2B2*T4/
        SUB=Case  DDFM=sat;
        REPEATED / sub=Case  group=Phase type=ar(1);
        ODS OUTPUT solutionF=fixed4 covparms=random4 fitstatistics=fit4;
RUN;
```

Model 4 is similar to model 3 with the only change that a fixed predictor, Class, is added in the model-option.

5. Logistic Model

Logistic Model A

```
PROC GLIMMIX  DATA=TwoLevel;
        CLASS Case ;
        MODEL Y/10 =  Phase   / SOLUTION DIST=binomial LINK=logit CL;
        RANDOM Intercept Phase /SUB=Case;
RUN;
```

Logistic Model B

```
PROC GLIMMIX  DATA=TwoLevel;
        CLASS Case ;
        MODEL Y/10 =  A1B1 B1A2 A2B2  / SOLUTION DIST=binomial LINK=logit CL;
        RANDOM Intercept A1B1 B1A2 A2B2  /SUB=Case;
        ODS OUTPUT solutionF=fixed5 covparms=random5 fitstatistics=fit5;
RUN;
```

For count data, the glimmix procedure is called. The other options are similar as in Model 1, with the only difference that the type of distribution has to be defined using the dist-option and the link-option. The dependent variable, Y, is divided by 10 because the outcome variable is a count out of ten.

Addendum A6: SAS code three-level analysis (Chapter 7)

Model 1: Basic Three-Level Model

```
PROC MIXED covtest DATA=ThreeLevel METHOD=ML;
        CLASS Study Case;
        MODEL Y = Phase / SOLUTION DDFM=sat;
        RANDOM Intercept  Phase / SUB=Study;
        RANDOM Intercept  Phase / SUB=Case(Study);
RUN;
```

The code for the three-level modeling is similar to the one used for the two-level modeling (see Appendix A). The only difference is an additional categorical variable, namely Study, defined in the class statement. We also have an additional random statement to indicate that the intercept and phase randomly vary across cases and across studies. The modeling of autocorrelation, heterogeneous within-case variance (Model 2), linear trends (Model 3) and predictor at the second level (Model 4) is similar as in the two-level modeling.

Addendum B: Raw data

Raw data Multiple-baseline data for the first two participants of the study of Laski et al. (1988) (Chapter 6)

Case	Time	Time1	Time2	Treatment	Treatment*Time	Treatment*Time1	Treatment*Time2	Y
1	0	-4	0	0	0	0	0	27.60
1	1	-3	1	0	0	0	0	23.96
1	2	-2	2	0	0	0	0	23.83
1	3	-1	3	0	0	0	0	47.26
1	4	0	4	1	4	0	4	52.70
1	5	1	4	1	5	1	4	60.99
1	6	2	4	1	6	2	4	66.6
1	7	3	4	1	7	3	4	52.50
1	8	4	4	1	8	4	4	85.88
1	9	5	4	1	9	5	4	47.05
1	10	6	4	1	10	6	4	66.28
1	11	7	4	1	11	7	4	54.05
1	12	8	4	1	12	8	4	51.23
2	0	-5	0	0	0	0	0	49.67
2	1	-4	1	0	0	0	0	23.57
2	2	-3	2	0	0	0	0	26.38
2	3	-2	3	0	0	0	0	28.35
2	4	-1	4	0	0	0	0	45.11
2	5	0	5	1	5	0	5	70.67
2	6	1	5	1	6	1	5	79.13
2	7	2	5	1	7	2	5	84.09
2	8	3	5	1	8	3	5	88.56
2	9	4	5	1	9	4	5	80.90
2	10	5	5	1	10	5	5	91.84
2	11	6	5	1	11	6	5	63.42
2	12	7	5	1	12	7	5	70.38